EVIC-VL: EVIDENCE-CONDITIONED VERIFICATION FOR CONFIDENT OBJECT HALLUCINATIONS IN VISION–LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision–language models (VLMs) can produce object claims that are both linguistically confident and unsupported by the image. Such claims can evade verification policies that intervene only when a model exposes high self-uncertainty. We study this failure mode as *confident visual hallucination* (CVH). We operationalize CVH at the object-claim level and construct CONFHALL-VL, a COCO-derived diagnostic that labels each candidate purely from the caption's own confidence and uncertainty signals together with the image annotations, without any detector evidence. We then propose EVIC-VL, an evidence-conditioned procedure that queries two frozen open-vocabulary object detectors, OWLv2 and Grounding DINO, with canonical object names, averages their evidence scores, and applies a validation-frozen policy to the original caption. As a secondary robustness check, EVIC-VL matches the frozen base generator on standard hallucination benchmarks overall, with its largest gain on object-presence polling accuracy (POPE F1).Our central question is whether combining these independent detectors improves object-level CVH discrimination over a single detector. On a sealed holdout, the equal-weight ensemble reliably raises the area under the ROC curve (AUROC) over the single-detector baseline by about four points.Overall, EVIC-VL treats detector output as fallible evidence rather than ground truth, and this design measurably improves confident-hallucination detection.

1 INTRODUCTION

Large vision–language models (VLMs) can describe objects that are not present in an image, even when the generated text is fluent and plausible (Rohrbach et al., 2018; Li et al., 2023c). Modern VLMs combine contrastively pre-trained image–text encoders (Radford et al., 2021) with instruction-tuned language backbones (Alayrac et al., 2022; Li et al., 2023a; Dai et al., 2023; Liu et al., 2023; Zhu et al., 2024). Neural-network confidence scores are known to be miscalibrated in general (Guo et al., 2017), so a fluent, high-probability claim need not be a supported one. Existing evaluation tools quantify object hallucination through caption overlap or object-presence questions, while mitigation methods modify decoding, train hallucination-aware models, or revise outputs with external evidence. These tools do not isolate one operational hazard. A generator can mention an object that is not in the image while assigning it high token probability and emitting little self-uncertainty.

We call such a claim a *confident visual hallucination* (CVH). CVH is a property of recorded artifacts, not of model internals: a claim is CVH when the model assigns it high token probability, emits little self-uncertainty, and its object is absent from the image's annotations. Nothing else enters the label, and we restrict claims to canonical COCO objects. A generator can also be trained to catch such cases itself (Wu et al., 2025); Sec. 2 positions that lineage and our split from it. In contrast, we keep the generator fixed and verify recorded claims with external visual evidence. Our narrow question is whether this evidence finds confident object claims that uncertainty alone ranks poorly, and whether acting on it improves a caption without erasing supported content.

Research question. We fix the generated claims, the image set, the mapping of claims to canonical objects, and the evaluation unit; the unit is one canonical object within one image. A detector's evidence score for a claim is its confidence that the claimed object is present in the image. Our

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.

primary question compares two scores on a prospective holdout, whose images were fixed before any evaluation of them: the OWLv2 evidence score alone, and the pre-specified equal-weight combination of the OWLv2 and Grounding DINO scores, from two frozen open-vocabulary detectors. Does the combined score separate CVH object claims better than the OWLv2 score alone? We answer in the affirmative: in the area under the receiver operating characteristic curve, AUROC hereafter, the combination reaches 0.927 and OWLv2 alone 0.887, a paired gain of +0.0401 with a 95% interval [+0.0289, +0.0515] from resampling whole images (Sec. 7), passing both pre-registered gates. A second question asks whether edits that deassert flagged claims in the same caption improve Semantic CVH Edit F1, hereafter *Edit F1*, a per-claim score of whether CVH claims are removed and supported claims kept, computed from ratings by raters who never see method names. We execute that rating protocol as a blinded 105-item human study (Sec. 6.2, App. T): the two editors attain near-identical Edit F1 while EVIC-VL leaves far more supported claims untouched (1/24 vs. 9/24 control-claim interventions). The detector-score AUROC measures fixed-claim detection only and is not evidence of editing effectiveness.

Contributions. We make three contributions:

- **An object-level CVH diagnostic.** We formalize CVH from three recorded signals: the generator's confidence, its self-uncertainty, and the absence of the object in the image annotations. We then construct CONFHALL-VL, a diagnostic set that keeps every claim occurrence linked to its canonical object, stores the exact editor input, and ships with provenance checks.
- **Pre-specified evidence-conditioned verification.** Our method EVIC-VL maps each object claim to a canonical query, combines the two frozen detectors' evidence scores with equal weights, and flags a claim when its risk exceeds a threshold chosen on validation only. It updates no parameters, but it does rely on pretrained VLM and detector checkpoints. Detector scores enter as fallible evidence, never as ground truth.
- **Evaluation separating detection, editing, and generation.** We pre-specify paired image-cluster inference, release the editing metric and blind rating protocol of Sec. 6.2 — T3; the registered tracks are enumerated in Table 17, App. R — and execute a stratified blinded audit of CONFHALL-VL labels with control probes and a washout re-rating. The audit finds a CVH-candidate precision of 0.918, an inter-rater Cohen's κ of 0.497, and a self-agreement of 0.925 on the washout re-rating (App. Y). We keep fixed-claim detection, same-caption editing, and independently generated captions separate, and we prohibit cross-caption rebasing.

Scope and claim boundary. We claim the definitions, the implemented procedures, the pre-specified evaluation design, the completed prospective detector contrast (Sec. 7), and the executed label-audit measurements (App. Y), and nothing more. We do not claim that EVIC-VL beats any baseline on editing or cross-model axes. CONFHALL-VL is English and COCO-derived; it does not evaluate attributes, relations, counting, text in images, actions, or scenes. When a generator exposes no uncertainty value, we treat its self-uncertainty as zero; this is a protocol convention, not a measured uncertainty. We do not claim that any score here is a calibrated probability, that EVIC-VL mitigates hallucinations universally, or that the coverage diagnostic enforces content preservation.

2 RELATED WORK

Measuring object hallucination. Object hallucination has been measured along two complementary axes: whether caption-mentioned objects can be matched to image annotations, and whether an evaluator can judge object faithfulness from the output itself. On the first axis, CHAIR compares objects mentioned in a caption with image annotations (Rohrbach et al., 2018), and OpenCHAIR lifts that discipline to open vocabularies (Ben-Kish et al., 2024); classic caption metrics measure consensus and semantic content rather than object faithfulness (Vedantam et al., 2015; Anderson et al., 2016), and correlate poorly with object hallucination (Rohrbach et al., 2018). On the second axis, POPE evaluates object presence through controlled polling questions (Li et al., 2023c), AMBER broadens generative and discriminative hallucination evaluation (Wang et al., 2024), MMHal-Bench grades open-ended responses with GPT-4 (Sun et al., 2024), HallusionBench separates language

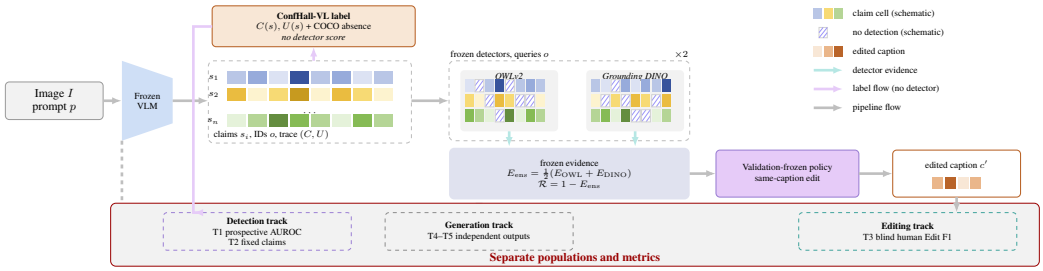

Figure 1: Registered EVIC-VL workflow. A frozen VLM produces a base caption and claim traces. CONFHALL-VL labels initial CVH candidates without using inference-time detector scores. EVIC-VL obtains canonical-object evidence from frozen detectors, computes a fixed risk score, and edits the same base caption. Detection (T1–T2) and independent generation (T4–T5) are distinct measured tracks; same-caption editing (T3) is the measured track of the executed blinded study built on the frozen protocol and metric of Sec. 6.2.

priors from genuine visual illusion (Guan et al., 2024), FaithScore verifies atomic facts in free-form outputs against the image (Jing et al., 2024), M-HalDetect targets detection directly with fine-grained instance annotations (Gunjal et al., 2024), and GAVIE evaluates instruction following without gold answers (Liu et al., 2024a). These benchmarks give useful task-level measurements but do not isolate the event we target: reading each fixed object claim's generation trace, we ask whether a claim produced with high confidence and low uncertainty lacks image support. We still report the standard benchmarks, as checks that our method does not degrade standard hallucination behavior.

Hallucination mitigation. Mitigation methods intervene at different points: at decoding, during alignment training, or after generation completes. At decoding, VCD contrasts token distributions under original and distorted visual inputs (Leng et al., 2024), following contrastive decoding for text-only models (Li et al., 2023b; Chuang et al., 2024); OPERA penalizes over-trusted tokens with retrospection (Huang et al., 2024); HALC couples focal visual grounding with contrastive decoding (Chen et al., 2024); and SID introspects self-attention token importance and subtracts amplified hallucination logits during decoding (Huo et al., 2025). On the alignment axis, RLHF-V learns from fine-grained human corrections (Yu et al., 2024), factually augmented reward modeling counteracts reward hacking (Sun et al., 2024), preference optimization targets hallucination directly (Zhao et al., 2023; Yang et al., 2025), calibrated self-rewarding removes the external judge (Zhou et al., 2024b), and hallucination-augmented contrastive learning separates faithful from hallucinated captions (Jiang et al., 2024). Post-hoc systems repair finished captions: Woodpecker corrects with visual knowledge (Yin et al., 2024), LURE learns a dedicated revisor (Zhou et al., 2024a), and Volcano revises via self-feedback (Lee et al., 2024). REVERSE instead trains a VLM to expose verification states and retrospectively resample (Wu et al., 2025). We follow this lineage of exposing verification signals, but split the problem into diagnosis and intervention: rather than modifying generation itself, we label fixed claims and trigger same-caption edits only after measurement.

Open-vocabulary evidence. Open-vocabulary detection grew from distillation, grounded pre-training, and image-level supervision (Gu et al., 2022; Li et al., 2022; Zhou et al., 2022) into transformer-based detectors: OWL-ViT transfers contrastive encoders to detection (Minderer et al., 2022), OWLv2 scales open-vocabulary detection through self-training (Minderer et al., 2023), and Grounding DINO couples grounded pre-training with open-set detection (Liu et al., 2024c); YOLO-World realizes real-time open-vocabulary detection (Cheng et al., 2024). A single registered measurement anchors this design: the pre-specified equal-weight combination outranks the stronger single-detector baseline. We claim the measurement, not the mechanism.

3 CONFIDENT VISUAL HALLUCINATION

Let I be an image, x a prompt, and $y = (y_1, \ldots, y_T)$ a caption produced by a frozen VLM. An object mention in the caption is a span s, covering token indices T_s. A deterministic synonym table maps each span to zero or more canonical COCO objects. We restrict the benchmark population to spans

with exactly one such target object, written $o = g(s)$. The main symbols are collected in App. A. Throughout, "frozen" marks anything fixed before evaluation and unchanged since, including model and detector checkpoints, caches, thresholds, policies, and artifacts.

3.1 Generation Confidence and Self-Uncertainty

Every generated token carries the model's own probability at decode time. We define length-normalized generation confidence as the geometric mean of these recorded probabilities over the span,

$$C(s) = \exp\left(\frac{1}{|T_s|}\sum_{u\in T_s}\log p(y_u \mid y_{<u}, I, x)\right). \tag{1}$$

$U(s)$ is the generation trace's recorded self-uncertainty value, if one is recorded. In REVERSE-style generators, this is the recorded probability of the unconfident closing state </UN> (Wu et al., 2025). For generators whose trace records none, the construction instead records $U(s) = 0$ by convention and marks the label source accordingly. Self-reported certainty is informative yet imperfect (Kadavath et al., 2022). Consequently, analyses that interpret low U as measured self-certainty must be restricted to traces where U was observed.

3.2 Operational CVH Label

Let $A_I \subseteq \mathcal{O}_{\mathrm{COCO}}$ be the set of COCO objects associated with image I: the union of the instance annotations and the object words extracted from the image's reference captions, following the CHAIR ground-truth construction (Rohrbach et al., 2018). For a canonical object claim $o = g(s)$, the initial label is

$$z(s, I) = \mathbb{I}[C(s) \geq 0.80]\ \mathbb{I}[U(s) \leq 0.15]\ \mathbb{I}[o \in \mathcal{O}_{\mathrm{COCO}}]\ \mathbb{I}[o \notin A_I]\,. \tag{2}$$

The rule reads recorded confidence and uncertainty from the generation trace and checks the object against the image's object set. For the registered confirmatory population, no OWLv2, Grounding DINO, combined, or EVIC-VL score enters Equation (2): this separation prevents the method from defining its own positive class. Annotation incompleteness remains a threat, quantified by the blinded label audit in App. Y.

3.3 ConfHall-VL Construction

The frozen construction begins with captions generated for COCO val2014 images using a fixed prompt and model revision (Lin et al., 2014; Chen et al., 2015; Liu et al., 2023; 2024b). It stores the unmodified base caption, not a concatenation of its spans, together with a SHA-256 digest of the editor input, per-occurrence claim IDs, canonical object IDs, C, U, and a status recording why each label decision was made. Repeated mentions retain separate occurrence identities; the detection evaluation later collapses them to one (I, o) row so that repeated nouns are not overweighted, and mixed-label groups are dropped and reported (App. C). Unresolved image IDs, absent annotations, and inconsistent derived fields fail closed: the affected item is excluded and the discrepancy surfaces as an error, never as a guessed value.

Audit independence. Grounding DINO is part of EVIC-VL, so DINO-based automatic demotion, relabeling a CVH candidate when the detector finds its object, cannot by itself establish method-independent confirmatory labels. Detector output may triage candidates for blinded review or support a sensitivity analysis, but the annotation-defined labels remain primary; label quality is established by a blinded human audit that does not use method outputs (App. Y).

4 EVIC-VL

4.1 Canonical-Object Evidence

Two frozen open-vocabulary detectors score each canonical object claim. For detector $d \in \{\mathrm{OWL}, \mathrm{DINO}\}$, let $\mathcal{B}_d(o, I)$ be its boxes for canonical query o and let $q_d(b, o, I) \in [0, 1]$ be the

detector's score for box b. The evidence for a claim is the best box score for that query,

$$E_d(o, I) = \max_{b \in \mathcal{B}_d(o,I)} q_d(b, o, I), \qquad \max \varnothing = 0. \tag{3}$$

Detector outputs are computed once and stored in read-only caches. A missing cache entry is not an empty detection set: the claim is *not checked*, and the run fails closed. The confirmatory score is the equal-weight combination, which we call the ensemble,

$$E_{\text{ens}}(o, I) = \tfrac{1}{2} \left(E_{\text{OWL}}(o, I) + E_{\text{DINO}}(o, I) \right). \tag{4}$$

No detector selection or weighting is performed on the holdout.

4.2 Risk and Intervention

The single-detector and combined ranking scores are

$$\mathcal{R}_{\text{OWL}} = 1 - E_{\text{OWL}}, \qquad \mathcal{R}_{\text{EViC}} = 1 - E_{\text{ens}}. \tag{5}$$

The registered detection contrast is threshold-free; it ranks claims without fixing an operating point. For same-caption intervention, validation fixes t=0.73 before the holdout is constructed; the policy flags an object claim when $\mathcal{R}_{\text{EViC}} \geq t$. This threshold maximizes the F1 of CVH detection on the development split subject to leaving at least 0.90 of supported claims unflagged. These validation figures select an operating point and are not confirmatory performance estimates.[1]

For editing, the editor receives the exact stored base caption and chooses one of five per-claim actions: keep the claim; drop it; hedge it to a possibility; replace it with a grounded alternative where one exists; or abstain on that claim alone. Each choice is logged. Claim-occurrence identity comes from the benchmark records of Sec. 3, not from the editor's text. A coverage diagnostic reports how much supported content the edit dropped; it does not override the editor and does not guarantee any minimum final coverage. In the metric terms of Sec. 6.2, a hedged claim is no longer positively asserted; it counts as deassertion, and neither hedging nor abstention counts as strict correction. The final guard against trivial delete-all and hedge-all editors is the Edit F1 floor of Proposition F5 together with the human metric and rating protocol of Sec. 6.2. The executed blinded human study is reported in Sec. 7 and App. T. The complete inference-time contract, from detector scoring to this edit path, is written out as Algorithm 1 in App. B.

The code also supports

$$\mathcal{R}_{\alpha,\beta,\gamma}(s, I) = \alpha U(s) + \beta(1 - E(s, I)) + \gamma[C(s) - E(s, I)]_+, \tag{6}$$

where $[\cdot]_+$ denotes the positive part and $E(s, I)$ is the ensemble evidence of the span's canonical object. This family is useful for validation ablations. It is not the frozen confirmatory candidate. On the development (dev) split we grid the weight simplex, $\alpha, \beta, \gamma \geq 0$ and $\alpha + \beta + \gamma = 1$, at step 0.05 (231 points). Every point on the evidence-only edge (γ=0) with $\beta > 0$ ties at the dev optimum; only the uncertainty-only corner $(1, 0, 0)$ degenerates. The recorded best point is exactly the frozen equal-weight ranking (α, β, γ)=$(0, 1, 0)$, with dev AUROC 0.9238 (App. L). We attribute no test or holdout result to this family. Setting $(\alpha, \beta, \gamma) = (1, 0, 0)$ recovers only an uncertainty score, not REVERSE's training, backtracking, or resampling.

5 Evaluation Setup

This section states the data, the processing, how experiments run, and the models; all outcomes use the metrics defined in Sec. 6.

5.1 Data

We evaluate on the datasets in Table 1. ConfHall-VL is built from COCO val2014 (Lin et al., 2014; Chen et al., 2015). The existing 1,300 development/test captions participated in candidate selection

[1]The frozen point stands as the prospectively frozen choice: re-running the same selection rule on the junction-fixed development build — that is, the build with the corrected claim–evidence join used for the reported cells selects a slightly lower operating point.

Table 1: Evaluation datasets. ConfHall-VL rows follow the T1-protocol LLaVA-1.5-7B column, and per-column populations appear in Table 7, App. D. The holdout split is the prospective confirmatory population, while the dev and test splits remain validation or exploratory. POPE (Li et al., 2023c) aggregates its random, popular, and adversarial settings over the same 500 images, and together with AMBER (Wang et al., 2024) forms a separate context track outside the confirmatory population.

Dataset	Role	Images	Object claims	CVH	Source
ConfHall-VL dev300	validation	300	446	53	COCO val2014
ConfHall-VL test1000	validation	1,000	1,544	188	COCO val2014
ConfHall-VL holdout2100	confirmatory	2,100	3,006	298	COCO val2014
POPE, three settings	T5 context	500	9,000 questions		COCO val2014
AMBER generative	T5 context	1,004	1,004 captions		external benchmark

Table 2: Frozen backbones. Each backbone is pinned to one public revision held constant across every reported run, with revisions recorded in the identity manifest specified in App. R. LLaVA-1.5-7B is the only T1-protocol column and its labels are measured on caption token streams. The four detector-cache columns carry restored labels and are not interchangeable evidence; provenance and per-column populations are in App. D.

Backbone	Scale
LLaVA-1.5-7B	7B
Qwen3-VL-8B-Instruct	8B
gemma-4-31B-it	31B
Qwen3-VL-32B-Instruct	32B
Qwen3-VL-235B-A22B-Instruct	235B

and are validation or exploratory data; selection on such data biases evaluation upward (Cawley & Talbot, 2010; Varma & Simon, 2006), so the confirmatory population is 2,100 previously unused images. To keep it untouched and unbiased, we exclude every ID and content hash used in earlier pilot, smoke, development, test, raw-output, or cache artifacts and rank the eligible IDs by a seeded pseudo-random order that is deterministic and reads neither images nor outcomes. The selection procedure is pinned in App. R.

5.2 Protocol Binding

Experiments run under a protocol that fixes identifiers and hashes, labels, caches, revisions, code, configuration, the endpoint, bootstrap settings, and input checksums before evaluation (App. R). The append-only ledger stores the first protocol fingerprint, the content hash of the frozen protocol: resuming requires an exact match; anything else becomes exploratory. A second person reviewed the frozen protocol before its one-time opening, and the binding happened exactly once under the recorded fingerprint; Sec. 7 reports the outcome.

5.3 Models

Five frozen backbones span families, generations, and scale (7B–235B; Table 2), covering the measurement problem rather than method-generalization claims.

6 Evaluation Metrics

6.1 Metric Digest

We fix here how outcomes are scored and judged. Table 3 lists in plain terms every metric that carries a reported value in the main text, together with the released editing metric (Edit F1) whose blinded human study is executed (App. T); the formal definitions and conventions are collected in App. U. The eight registered tracks T1–T8, with their questions, populations, headline metrics, and validity gates, are enumerated in Table 17 (App. R). Detection (T1–T2) and editing (T3) are scored only on

Table 3: Plain-language digest of the metrics reported or released in the main text; formal definitions and conventions are collected in App. U.

Metric	In words
AUROC	probability that a randomly drawn CVH claim scores above a randomly drawn supported claim
AUPRC, the area under the precision–recall curve	non-interpolated average precision across thresholds
ΔAUROC	paired contrast, ensemble vs. OWLv2-only score
percentile CI	image-cluster resampled interval (3,000 replicates, seed 0)
BCa CI	bias-corrected and accelerated interval of the same replicates
CVH recall	share of CVH claims flagged at t=0.73
supported specificity	share of supported claims not flagged at t=0.73
median three-term risk	label-stratified median of the legacy validation risk on the fixed claims (context)
Edit F1	per-claim score of whether CVH claims are removed and supported claims kept (Eq. (8)); T3 headline, executed blinded human study (App. T)
residual risk $\hat{R}$	share of ground-truth-absent post-edit mentions still passing the calibrated τ
revision rate	share of captions the editing protocol revises
CHAIR, context metric	caption-mentioned objects vs. the image's reference annotations
CVH precision	audited retained-stratum candidates confirmed absent, excluding unresolvable cases
control false-alarm	guaranteed-present controls judged absent
Cohen's κ	chance-corrected inter-rater agreement
self-agreement	same-rater repeat rate on the blinded washout re-rating
second-rater absent rate	second rater's absent share on the shared triage-retained rows

the claims of the stored base captions; the independent generations of T4–T5 are scored separately and are never re-entered there as if they were edits.

6.2 Primary Endpoint and the Editing Metric

The unit is a canonical object within an image. Let z_{io} be its CVH label. The pre-specified contrast is

$$\widehat{\Delta}_{\mathrm{AUROC}} = \mathrm{AUROC}(\mathcal{R}_{\mathrm{EViC}}, z) - \mathrm{AUROC}(\mathcal{R}_{\mathrm{OWL}}, z). \tag{7}$$

We form a paired percentile confidence interval (CI) by resampling images with replacement and including all object rows from each sampled image (3,000 replicates, seed 0, $\alpha = 0.05$). Statistical superiority requires the 95% interval lower bound to exceed zero. Scientific relevance is reported separately against a smallest effect size of interest (SESOI) of 0.02 AUROC. The headline claim requires both gates; neither substitutes for the other. The area under the precision–recall curve (AUPRC) and operating-point metrics are secondary.

Each editor receives the byte-identical base caption. At the object-claim level, a CVH claim no longer positively asserted is a true positive (TP), a CVH claim still asserted is a false negative (FN), and a supported claim no longer asserted is a false positive (FP). We define

$$\mathrm{EditF1} = \frac{2\mathrm{TP}}{2\mathrm{TP} + \mathrm{FP} + \mathrm{FN}}. \tag{8}$$

Hedging counts as deassertion but not strict correction. Full-caption abstention deasserts all CVH claims but loses all supported claims. The headline metric is accompanied by secondary descriptors—deassertion recall, supported assertion retention, strict correction rate, new unsupported object-claim rate, evaluation coverage, and abstention rate—defined in App. U (Eq. (18)).

6.3 Released Rating Protocol

The released protocol specifies two raters under method-blind identifiers with an adjudicator resolving every disagreement (full rubric: App. T); inter-rater agreement and Cohen's κ are to be reported (Cohen, 1960). The automatic semantic extractor is a development aid and secondary analysis, never the T3 headline (Table 17). This submission executes the blinded human study under the released metric and protocol; the results are reported in Sec. 7 and App. T.

Table 4: T1: prospective object-level detection. Intervals are paired by image cluster (Cameron & Miller, 2015). The registered claim passes only when the lower bound of the interval exceeds zero and the point gain reaches 0.02. As a cross-check, the BCa interval from the same 3,000-replicate bootstrap (Efron, 1987; Efron & Tibshirani, 1993) is [+0.0291, +0.0515], closely tracking the reported percentile interval. The Grounding DINO row separates true complementarity from weak-noise averaging; the ΔAUROC column reports the registered paired contrast against OWLv2 as the reference row, and the Grounding DINO row has no registered contrast and displays –. Bold marks the best cell per column; all ties are bold. Registered as T1 in Table 17. Metric definitions are collected in App. U.

Method	AUROC ↑	AUPRC ↑	ΔAUROC [95% CI]
OWLv2	0.887	0.478	reference
Grounding DINO	0.850	0.462	–
OWLv2 + DINO (equal weight)	**0.927**	**0.623**	**+0.0401 [+0.0289, +0.0515]**
Population: 3,006 object claims; 298 CVH claims.			

Table 5: T2: mechanism analysis on identical fixed claims at the frozen operating point $t{=}0.73$ of Sec. 4. The U-only row is a scalar uncertainty baseline and not the complete REVERSE algorithm. Recall is measured on CVH claims and specificity on supported object claims. The three-term risk row keeps the legacy validation-family rule for lineage and is excluded from the method claims; the frozen policy's flag rule coincides with the thresholded OWLv2+DINO score row. Bold marks the best cell per column; all ties are bold. Registered as T2 in Table 17. Metric definitions are collected in App. U.

Score / policy	CVH recall ↑	Supported specificity ↑
U-only	0.000	**1.000**
OWLv2 only	**0.802**	0.800
OWLv2 + DINO score	0.614	0.954
three-term risk (legacy validation family)	0.674	0.894

7 Results

7.1 Confirmatory Detection

On the frozen prospective holdout (2,100 selected images; 1,837 contribute 3,006 evaluated object claims), the equal-weight ensemble reaches object AUROC 0.927 against 0.887 for the single-detector OWLv2 baseline (Table 4): a paired gain of +0.0401 with 95% image-cluster CI [+0.0289, +0.0515]. We closed the prospective-holdout protocol before any holdout evaluation was run and recorded the closure in a gate record; App. R specifies its contents.

7.2 Mechanism on Fixed Claims

Table 5 reports the mechanism analysis on identical fixed claims at the frozen operating point. As context, the label-stratified median three-term risk on this population is 0.554 on CVH claims and 0.296 on supported claims; it is identical across score rows by construction.

7.3 External Guarantee Stress Test

Conformal Evidence-Bounded Control (CEBC) (Mishra et al., 2026) edits captions under a pre-registered, conformally calibrated residual-risk guarantee, with calibration on a split disjoint from evaluation. We re-implement its Table 1 setup on eval500, the 500-image evaluation split, with dual ground-truth calibration arms P and C, with frozen constants, and ask whether the nominal guarantee survives measured label noise.

At the calibrated $\tau{=}0.7197$, identical on both ground-truth arms, all three edit arms — arm A: calib-P/eval-P, arm B: calib-P/eval-C, arm C: calib-C/eval-C, per App. V — hit the pre-registered break rule: the lower CI bound of the residual risk $\hat{R}$ — the share of ground-truth-absent post-edit mentions

whose evidence still passes τ — exceeds δ=0.05. The revision rate under our re-implementation is 94.2%, versus 31.2% in the original system; the residual-risk point estimate is 1.000 with a degenerate zero-width CI; the CHAIR side metrics CHAIR-S/CHAIR-I are $0.008/0.005$ on the edited captions. Full details, including the implementation-level gap in revision rate, are disclosed in App. V. "BROKEN" here is the pre-registered output label of this external stress test, not a verdict on EVIC-VL's registered detection contrast.

The same-caption editing policy at the validation-frozen operating point is released together with the Semantic CVH Edit F1 definition and the method-blind rating protocol; the blinded human study is executed under that protocol and metric (App. T); the executed outcome is reported in the Conclusion.

The generation, standard-benchmark, and cross-model tables appear in App. X; the label-audit and sensitivity tables are collected in App. Y and App. L. Every registered empirical cell is populated from a frozen artifact or from the executed blinded audit; the manuscript build fails closed whenever a registered cell references an unpopulated macro.

Visual variants of the registered tables and the sensitivity analyses are collected in App. P.

The frozen design carries five formal guards: equal-weight averaging of evidence never inflates score variance beyond the component average; the paired interval weakly dominates independent intervals when the two evidence scores covary non-negatively on shared claims; the Wilson interval used for audit proportions stays within $[0, 1]$ with pinned edges; the audit-based correction of label-induced prevalence is unbiased under one-sided noise, constant audit precision, and an unbiased audited subsample of the flagged claims; and a trivial delete-all editor is floored at $\mathrm{EditF1}_{\varnothing} = 2\pi/(1+\pi)$, dominated once $r > \pi + (1-\pi)\,q$. Statements are collected as propositions in App. F with assumptions and full proofs in App. G.

8 LIMITATIONS

The benchmark addresses English COCO object claims only, not attributes, relations, counts, in-image text, actions, scenes, or open-world factuality, and for generators whose traces record no self-uncertainty the uncertainty-conditioned readings are unavailable. The labels are checked rather than trusted through the audit of App. Y: the author rater judged all 795 rows and a second blinded rater 160 of them, with residual bias not ruled out at Cohen's κ of 0.497 and the control false-alarm share 0.182. Detector scores are fallible evidence pinned to the frozen revisions, not calibrated probabilities. Editing trades content for fewer hallucinations, and since neither editor in the executed blinded study clears the prevalence-dependent Edit F1 floor of Proposition F5 and their intervals overlap, we report the Edit F1–retention pair and claim no editing advantage. Build completeness is not scientific completeness: the T1 holdout gate ran once under the registered protocol, the T3 cells come from the executed study, and the cross-model editing question (T6) is out of scope.

9 CONCLUSION

We formulate confident object hallucination as a fixed-claim disagreement between a generator trace and image support, specify a provenance-preserving COCO-derived diagnostic, and introduce an evidence-conditioned detector ensemble with a same-caption intervention protocol. The evaluation design makes the primary detector contrast prospective, keeps editing separate from independent generation, and penalizes destructive correction through supported claim loss. The prospective holdout contrast is positive (paired ΔAUROC $+0.0401$) and passes both pre-specified gates, and the accompanying audit estimates the label quality of the diagnostic. In an executed 105-item method-blind editing study, EVIC-VL attains Edit F1 0.7792 versus 0.7500 for detector revision, with overlapping intervals and the contrast carried by supported-claim preservation (App. T); cross-model editing replication remains an open scientific question.

AI USE STATEMENT

Per the ICLR 2027 AI policy, we disclose that generative AI tools were used in the preparation of this work as follows: to help polish the writing of the manuscript, and to assist with implementation during

research execution, including code for the evaluation pipeline and the aggregation and analysis scripts. The research ideas, experimental design, evaluation, and interpretation of results were developed and determined by the authors. The authors made all final decisions regarding the structure and wording of this paper. All AI-assisted material was reviewed and revised by the authors, with automated checks used where helpful. We take full responsibility for the content of this paper, including text, claims, and artifacts.

Ethics Statement

Reducing unsupported visual claims could help systems used for captioning, search, and human decision support. The same mechanism can also create false assurance: detector agreement is not proof of presence, and abstention may hide unequal performance across object categories or domains. We therefore report coverage and category sensitivity alongside hallucination metrics, and we do not present EVIC-VL as a safety guarantee. The label audit is carried out under full blinding; an author serves as a rater, and the pre-specified protocol admits an "unresolvable" judgment (App. Y). No personal information was collected. All third-party images, annotations, and model checkpoints were used in compliance with their respective licenses and data terms.

Reproducibility Statement

Anonymized source code for the evaluation pipeline, together with frozen evaluation artifacts and an identity manifest recording the repository commit, the protocol fingerprint, and the pinned model and detector revisions behind every reported run, is provided as supplementary material. The evaluation design is specified and frozen prior to holdout evaluation; the protocol and its holdout-closure criteria are referenced in Sec. 5 and App. R. Metric definitions are collected in App. U; data processing steps, provenance, and exclusion accounting are described in App. S; the claim–evidence map is App. Z; and the formal guards are stated as propositions in App. F, with their assumptions and complete proofs in App. G; the guard lemmas of App. C are also proved there. The label-audit protocol, including blindness, control probes, and washout re-rating, is specified in App. Y. All reported empirical values derive only from frozen evaluation artifacts; mock, smoke-test, validation, and legacy artifacts are excluded from the manuscript.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikołaj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi:10.52202/068431-1723. URL https://proceedings.neurips.cc/paper_files/paper/2022/hash/960a172bc7fbf0177ccccbb411a7d800-Abstract-Conference.html.

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. SPICE: Semantic propositional image caption evaluation. In *Computer Vision – ECCV 2016*, pp. 382–398. Springer, 2016. doi:10.1007/978-3-319-46454-1_24.

Anastasios N. Angelopoulos and Stephen Bates. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023. doi:10.1561/2200000101.

Assaf Ben-Kish, Moran Yanuka, Morris Alper, Raja Giryes, and Hadar Averbuch-Elor. Mitigating open-vocabulary caption hallucinations. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 22680–22698. Association for Computational Linguistics, 2024. doi:10.18653/v1/2024.emnlp-main.1263. URL https://aclanthology.org/2024.emnlp-main.1263/.

A. Colin Cameron and Douglas L. Miller. A practitioner's guide to cluster-robust inference. *Journal of Human Resources*, 50(2):317–372, 2015. doi:10.3368/jhr.50.2.317.

Gavin C. Cawley and Nicola L. C. Talbot. On over-fitting in model selection and subsequent selection bias in performance evaluation. *Journal of Machine Learning Research*, 11:2079–2107, 2010. URL https://www.jmlr.org/papers/v11/cawley10a.html.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015. URL https://arxiv.org/abs/1504.00325.

Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. HALC: Object hallucination reduction via adaptive focal-contrast decoding. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 7824–7846. PMLR, 2024. URL https://proceedings.mlr.press/v235/chen24bi.html.

Tianheng Cheng, Lin Song, Yixiao Ge, Wenyu Liu, Xinggang Wang, and Ying Shan. YOLO-World: Real-time open-vocabulary object detection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16901–16911, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Cheng_YOLO-World_Real-Time_Open-Vocabulary_Object_Detection_CVPR_2024_paper.html.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Th6NyL07na.

Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46, 1960. doi:10.1177/001316446002000104.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N. Fung, and Steven C. Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi:10.52202/075280-2142. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/9a6a435e75419a836fe47ab6793623e6-Abstract-Conference.html.

Bradley Efron. Better bootstrap confidence intervals. *Journal of the American Statistical Association*, 82(397):171–185, 1987. doi:10.1080/01621459.1987.10478410.

Bradley Efron and Robert J. Tibshirani. *An Introduction to the Bootstrap*. Number 57 in Monographs on Statistics and Applied Probability. Chapman and Hall, 1993.

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

Xiuye Gu, Tsung-Yi Lin, Weicheng Kuo, and Yin Cui. Open-vocabulary object detection via vision and language knowledge distillation. In *International Conference on Learning Representations*, 2022. URL https://arxiv.org/abs/2104.13921.

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. doi:10.1109/CVPR52733.2024.01363.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18135–18143, 2024. doi:10.1609/aaai.v38i16.29771.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1321–1330. PMLR, 2017. URL https://proceedings.mlr.press/v70/guo17a.html.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Huang_OPERA_Alleviating_Hallucination_in_Multi-Modal_Large_Language_Models_via_Over-Trust_CVPR_2024_paper.html.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=rsZwwjYHuD.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27026–27036, 2024. doi:10.1109/CVPR52733.2024.02553.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. FaithScore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pp. 5042–5063. Association for Computational Linguistics, 2024. doi:10.18653/v1/2024.findings-emnlp.290. URL https://aclanthology.org/2024.findings-emnlp.290/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, Deep Ganguli, Danny Hernandez, Josh Jacobson, Jackson Kernion, Shauna Kravec, Liane Lovitt, Kamal Ndousse, Catherine Olsson, Sam Ringer, Dario Amodei, Tom Brown, Jack Clark, Nicholas Joseph, Ben Mann, Sam McCandlish, Chris Olah, and Jared Kaplan. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*, 2022. URL https://arxiv.org/abs/2207.05221.

Seongyun Lee, Sue Hyun Park, Yongrae Jo, and Minjoon Seo. Volcano: Mitigating multimodal hallucination through self-feedback guided revision. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 391–404. Association for Computational Linguistics, 2024. doi:10.18653/v1/2024.naacl-long.23. URL https://aclanthology.org/2024.naacl-long.23/.

Jing Lei, Max G'Sell, Alessandro Rinaldo, Ryan J. Tibshirani, and Larry Wasserman. Distribution-free predictive inference for regression. *Journal of the American Statistical Association*, 113(523): 1094–1111, 2018. doi:10.1080/01621459.2017.1307116.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Leng_Mitigating_Object_Hallucinations_in_Large_Vision-Language_Models_through_Visual_Contrastive_CVPR_2024_paper.html.

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. URL https://proceedings.mlr.press/v202/li23q.html.

Liunian Harold Li, Pengchuan Zhang, Haotian Zhang, Jianwei Yang, Chunyuan Li, Yiwu Zhong, Lijuan Wang, Lu Yuan, Lei Zhang, Jenq-Neng Hwang, Kai-Wei Chang, and Jianfeng Gao. Grounded language-image pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10965–10975,

2022. URL https://openaccess.thecvf.com/content/CVPR2022/html/Li_Grounded_Language-Image_Pre-Training_CVPR_2022_paper.html.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12286–12312, Toronto, Canada, 2023b. Association for Computational Linguistics. doi:10.18653/v1/2023.acl-long.687. URL https://aclanthology.org/2023.acl-long.687/.

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. Association for Computational Linguistics, 2023c. doi:10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

Zhuohang Li, Chao Yan, Nicholas J. Jackson, Wendi Cui, Bo Li, Jiaxin Zhang, and Bradley A. Malin. Towards statistical factuality guarantee for large vision-language models. *arXiv preprint arXiv:2502.20560*, 2025.

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

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=J44HfH4JCg.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi:10.52202/075280-1516. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/6dcf277ea32ce3288914faf369fe6de0-Abstract-Conference.html.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024b. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Improved_Baselines_with_Visual_Instruction_Tuning_CVPR_2024_paper.html.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding DINO: Marrying DINO with grounded pre-training for open-set object detection. In *European Conference on Computer Vision*, pp. 38–55, 2024c. URL https://www.ecva.net/papers/eccv_2024/papers_ECCV/html/6319_ECCV_2024_paper.php.

Matthias Minderer, Alexey Gritsenko, Austin Stone, Maxim Neumann, Dirk Weissenborn, Alexey Dosovitskiy, Aravindh Mahendran, Anurag Arnab, Mostafa Dehghani, Zhuoran Shen, Xiao Wang, Xiaohua Zhai, Thomas Kipf, and Neil Houlsby. Simple open-vocabulary object detection. In *Computer Vision – ECCV 2022*, pp. 728–755. Springer, 2022. doi:10.1007/978-3-031-20080-9_42.

Matthias Minderer, Alexey Gritsenko, and Neil Houlsby. Scaling open-vocabulary object detection. In *Advances in Neural Information Processing Systems*, volume 36, 2023. doi:10.52202/075280-3191. URL https://proceedings.neurips.cc/paper_files/paper/2023/hash/e6d58fc68c0f3c36ae6e0e64478a69c0-Abstract-Conference.html.

Ashish Mishra, Tarun Kumar, Arpit Shah, Suparna Bhattacharya, and Martin Foltin. CEBC: Conformal evidence-bounded control for low-hallucination vision–language generation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 46193–46206, San Diego, California, United States, 2026. Association for Computational Linguistics. doi:10.18653/v1/2026.acl-long.2142. URL https://aclanthology.org/2026.acl-long.2142/.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021. URL https://proceedings.mlr.press/v139/radford21a.html.

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. Association for Computational Linguistics, 2018. doi:10.18653/v1/D18-1437. URL https://aclanthology.org/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. Association for Computational Linguistics, 2024. doi:10.18653/v1/2024.findings-acl.775. URL https://aclanthology.org/2024.findings-acl.775/.

Sudhir Varma and Richard Simon. Bias in error estimation when using cross-validation for model selection. *BMC Bioinformatics*, 7(1):91, 2006. doi:10.1186/1471-2105-7-91.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. CIDEr: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 4566–4575. IEEE, 2015. doi:10.1109/CVPR.2015.7299087.

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*, 2024. URL https://arxiv.org/abs/2311.07397.

Edwin B. Wilson. Probable inference, the law of succession, and statistical inference. *Journal of the American Statistical Association*, 22(158):209–212, 1927. doi:10.1080/01621459.1927.10502953.

Tsung-Han Wu, Heekyung Lee, Jiaxin Ge, Joseph E. Gonzalez, Trevor Darrell, and David M. Chan. Generate, but verify: Reducing hallucination in vision-language models with retrospective resampling. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi:10.52202/085713-2206. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/5eff08bd064f0cdd92182cdf6fd06b99-Abstract-Conference.html.

Zhihe Yang, Xufang Luo, Dongqi Han, Yunjian Xu, and Dongsheng Li. Mitigating hallucinations in large vision-language models via DPO: On-policy data hold the key. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10610–10620, 2025. doi:10.1109/CVPR52734.2025.00992.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105, 2024. doi:10.1007/s11432-024-4251-x. URL https://doi.org/10.1007/s11432-024-4251-x.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. RLHF-V: Towards trustworthy MLLMs via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13807–13816, 2024. doi:10.1109/CVPR52733.2024.01310.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing LVLMs through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023. URL https://arxiv.org/abs/2311.16839.

Xingyi Zhou, Rohit Girdhar, Armand Joulin, Philipp Krähenbühl, and Ishan Misra. Detecting twenty-thousand classes using image-level supervision. In *Computer Vision – ECCV 2022*, pp. 350–368. Springer, 2022. doi:10.1007/978-3-031-20077-9_21.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. LURE: Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations*, 2024a. URL https://openreview.net/forum?id=oZDJKTlOUe.

Yiyang Zhou, Zhiyuan Fan, Dongjie Cheng, Sihan Yang, Zhaorun Chen, Chenhang Cui, Xiyao Wang, Yun Li, Linjun Zhang, and Huaxiu Yao. Calibrated self-rewarding vision language models. In *Advances in Neural Information Processing Systems*, volume 37, 2024b. URL https://proceedings.neurips.cc/paper_files/paper/2024/hash/5c20c00504e0c049ec2370d0cceaf3c4-Abstract-Conference.html.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=1tZbq88f27.

Table 6: Principal notation and frozen constants.

Symbol	Meaning
I	input image
x	text prompt
y / y^*	base / final caption of an image (Sec. 3; Eq. (17))
s	span (a text mention of an object)
T_s	token indices covered by span s
o	canonical COCO object of a span (via $g(s)$, Sec. 3)
$g(s)$	deterministic synonym map from span to canonical COCO objects (Sec. 3)
$\mathcal{O}_{\mathrm{COCO}}$	the canonical COCO object vocabulary
A_I	COCO object set of image I (instance annotations $\cup$ object words from reference captions; Sec. 3)
$\mathcal{B}_d(o, I)$	boxes detected by detector d for canonical query o in image I
$q_d(b, o, I)$	detector d's score for box b
$E_d(o, I)$	evidence score of canonical object o in image I from detector d (Eq. (3))
$E_{\mathrm{ens}}(o, I)$	equal-weight ensemble of the per-detector evidence scores (Eq. (4))
$\mathcal{R}$ / $\mathcal{R}_{\mathrm{EViC}}$	risk score ($1 - E_{\mathrm{ens}}$, Eq. (5), unless a different subscript is stated, e.g. $\mathcal{R}_{\mathrm{OWL}}$, $\mathcal{R}_{\alpha,\beta,\gamma}$)
t	frozen ensemble operating point, t=0.73 (Sec. 4)
τ	calibrated evidence threshold of the external CEBC stress test (τ=0.7197 on both ground-truth arms; Sec. 7, App. U)
δ	break margin of the pre-registered guarantee rule (App. C), δ=0.05 (also the $(1-\delta)$ quantile level for τ, App. U)
n_{cal}	calibration set size, n_{cal}=120 (App. U)
$\hat{R}$	residual risk after editing (Eq. (17))
Rev.	fraction of images whose final caption differs from the greedy base caption (Eq. (17))
C, U	confidence / self-uncertainty of the generating model (Eq. (1))
κ	inter-rater agreement (Cohen's kappa; App. U)
n_{obj}	number of object-level evaluation rows (App. D)
P, C	dual ground-truth calibration arms in the CEBC stress test (presence-only / clean with audit-restored objects; distinct from confidence C; Sec. 7)
α, β, γ	risk-family weights in $\mathcal{R}_{\alpha,\beta,\gamma}$ (Sec. 4; α is also reused as an interval error level, stated where used)
$\widehat{\Delta}_{\mathrm{AUROC}}$	paired difference of two methods' AUROCs (Eq. (7))
AUROC / AUPRC	area under the ROC / precision–recall curve on the frozen population
i, j	indices whose ranges are fixed locally by each equation (spans or claims within one response; i also indexes images, as in z_{io})
z_{io}	CVH label of canonical object o on image i (per-occurrence source)
$z(s, I)$	per-occurrence initial CVH label of span s (Eq. (2)); object-level view: z_{io}
$z^\star$	audit-confirmed truth of a claim (App. F)
CVH thresholds	confidence ≥ 0.80, uncertainty ≤ 0.15 (Eq. (2))
s_i	i-th span of a caption (Alg. 1)
n	sample or replicate count, stated where used (calibration size, per-table evaluation population, or binomial trials)

A Notation

Table 6 lists the main symbols and the frozen constants. Notation local to a single definition is introduced where it is used (e.g., App. U, App. G, App. F).

B Inference and Editing Algorithm

Algorithm 1 states the complete inference-time contract: detection at the frozen operating point and the same-caption edit path. Caches are read-only during inference, and no component is retrained anywhere in the pipeline. The edit path never asserts an unsupported mention: the only action that inserts text replaces a flagged claim with a grounded alternative where one exists, and every other action keeps, drops, hedges, or abstains on an existing claim. Each choice is logged. The CVH labels $z(s, I)$ of Eq. (2) arise at benchmark construction in App. S and never enter this path; the method does not define its own positive class.

Algorithm 1 EVIC-VL inference-time detection and same-caption editing. Constants are frozen in Table 6.

Require: image I, prompt x, and the frozen OWLv2 and DINO caches
1: $y \leftarrow$ generate the base caption for (I, x) from the frozen VLM
2: extract object-mention spans $(s_1, \ldots, s_m)$ from y by the frozen rule of App. S
3: **for all** s_i with exactly one canonical object $o_i = g(s_i)$ **do**
4: $\quad E_i \leftarrow E_{\text{ens}}(o_i, I)$ read from the caches
5: $\quad \mathcal{R}_{\text{EViC},i} \leftarrow 1 - E_i$
6: **end for**
7: leave every remaining claim *not checked*, never flag it, and exclude it from the detection population
8: flag each checked span s_i iff $\mathcal{R}_{\text{EViC},i} \geq t$ for the frozen point t=0.73
9: **edit**: resolve each flagged mention with one logged action, chosen from keep, drop, hedge, replace with a grounded alternative where one exists, or abstain
10: **return** the final caption y^* and the flagged spans

In the loop, E_i is the ensemble evidence of Eq. (4) and $\mathcal{R}_{\text{EViC},i}$ is the risk of Eq. (5). The risk of a span depends only on its canonical object and the image, so repeated mentions of one object share a single flag. The object-level reading of Eq. (9) coincides with the per-span view, and evaluation collapses repeated occurrences under the per-occurrence label invariance lemma of App. C. Validation selects the frozen point by the rule of Sec. 4, and App. L reports how the detection trade-off moves around it.

C GUARDS AND PROOFS

We only prove guard properties, not guarantees we do not claim. Full proofs appear in App. G. Lemmas C1–C3 below correspond to Lemmas G1–G3 of App. G.

Lemma C1 (Per-occurrence label invariance)**.** *Let occurrences* (s_i) *carry per-occurrence labels and let the collapse map group by* $(\text{image}, \text{canonical})$*. If all spans in a group share the same label, the object-level population is invariant to span-permutation; conflicting groups are removed entirely and reported. We call this drop rule the* conflict gate*.*

Lemma C2 (Calibrated quantile coverage)**.** *If calibration and evaluation rows are exchangeable and the threshold is the pre-registered quantile on the calibration tail, then the evaluation-side error-control probability is at least the target level, and exceeds it by at most* $1/(n_{\text{cal}} + 1)$ *under continuity (Lei et al., 2018; Angelopoulos & Bates, 2023).*

Lemma C3 (Verdict implication of the break rule)**.** *Under the pre-registered rule, if and only if the lower CI bound of residual risk* $\hat{R}$ *exceeds* δ *we say the nominal guarantee is broken; this statement depends only on the measured point and interval, not on decoding choices outside the frozen contract.*

D POPULATION STATISTICS PER COLUMN AND SPLIT

This section reports the ConfHall-VL evaluation populations behind every headline number over the five backbone columns. Rows are object-level: per-occurrence labels are collapsed to (image, object) with the conflict gate (App. C), and the counts below are the post-collapse populations that metrics are computed over. The four detector-cache blocks contribute all images in every split with one exception (gemma dev300: 297 of 300), unlike the LLaVA-1.5-7B T1 column (Tables 1 and 2), which is why their object counts are larger; their captions also carry about twice as many object claims per image (2.1–2.3$\times$ for the Qwen columns, 1.7$\times$ for gemma).

Two reading notes. First, n_{images} counts images that contribute at least one evaluated object claim, not the split pool. Second, CVH-rate comparisons across blocks are descriptive only — labels are column-native (measured C/U on LLaVA, detector-cache restoration elsewhere).

Table 7: Evaluation populations. Per-occurrence labels are collapsed to (image, object) pairs where applicable (App. C). Bold rows separate backbone blocks. n_{images} counts contributing images, not the split pool (holdout: 1,837 of 2,100; dev300: 262 of 300).

split	n_{images}	n_{obj}	CVH objects
LLaVA-1.5-7B			
dev300	262	446	53
test1000	897	1544	188
holdout2100	1837	3006	298
Qwen3-VL-32B-Instruct			
dev300	300	968	169
test1000	1000	3147	534
holdout2100	2100	6625	1152
Qwen3-VL-235B-A22B-Instruct			
dev300	300	1039	181
test1000	1000	3379	659
holdout2100	2100	7052	1342
Qwen3-VL-8B-Instruct			
dev300	300	1001	172
test1000	1000	3231	587
holdout2100	2100	6679	1142
gemma-4-31B-it			
dev300	297	790	89
test1000	1000	2590	277
holdout2100	2100	5136	447

E Frozen Figures from the Pre-Junction-Fix Builds

The pre-junction-fix figures (object-level AUROC with the paired interval, and the operating-point sweep) are superseded by the T1 measurements in Table 4 and the junction-fixed sweep in Table 9 and are removed from the print build; the bit-exact images remain archived in the frozen repository zips. The per-column population statistics are in Table 7.

F Structural Properties

The registered design of Secs. 3–7 rests on a few structural statements. All are guarantees of the construction itself; full proofs appear in App. G.

Proposition F1 (Ensemble variance contract)**.** *Let E_1, E_2 be evidence scores viewed as random variables over a claim population with variances σ_1^2, σ_2^2 and covariance c, and let $\bar{E} = (E_1 + E_2)/2$. Then $\mathrm{Var}(\bar{E}) = \frac{1}{4}(\sigma_1^2 + \sigma_2^2 + 2c) \le \frac{1}{2}(\sigma_1^2 + \sigma_2^2)$, with equality iff the scores are perfectly positively correlated and have equal variance. Equal-weight averaging therefore never inflates score variance beyond the component average, and any imperfect correlation strictly reduces it relative to that bound. Variance reduction is not an AUROC statement; discrimination is the registered empirical question of Sec. 6.*

Proposition F2 (Pairing can only improve contrast variance)**.** *For two scores s', s evaluated on identical claims, $\mathrm{Var}(s' - s) = \sigma'^2 + \sigma^2 - 2\mathrm{Cov}(s', s) \le \sigma'^2 + \sigma^2$ whenever $\mathrm{Cov}(s', s) \ge 0$. When the two methods' errors are positively correlated on shared claims, the paired contrast of Equation (7) is estimated at strictly lower variance than a difference of independent comparisons; this motivates the paired image-cluster interval rather than two independent intervals.*

Proposition F3 (Wilson interval form)**.** *For k successes in n Bernoulli trials, the Wilson interval of Equation (19) is contained in $[0, 1]$, its endpoints are strictly increasing in k, and the documented edge pins return the conservative endpoints $[0, \cdot]$ at $k{=}0$ and $[\cdot, 1]$ at $k{=}n$. Nominal $1{-}\alpha$ asymptotic coverage follows the standard pivotal argument (proof in App. G).*

Proposition F4 (One-sided audit correction)**.** *Let $z \in \{0, 1\}$ be the annotation-defined CVH label of Eq. (2), and let $z^\star \in \{0, 1\}$ denote the audit-confirmed truth of that claim. Let the audit measure*

the audit-confirmed precision $\rho = \Pr(z^\star{=}1 \mid z{=}1)$, constant over claims, on an unbiased audited subsample of $\{z{=}1\}$, with unconfirmed claims trusted ($z^\star{=}0$ whenever $z{=}0$). Then, with $\hat{\pi}_z$ the empirical share of claims with $z{=}1$ in the evaluation population, the corrected prevalence estimator $\hat{\pi}_{z^\star} = \hat{\rho} \cdot \hat{\pi}_z$ is unbiased for the true prevalence $\pi_{z^\star} = \Pr(z^\star{=}1)$, and per-claim expectations $\mathbb{E}[1\{z^\star{=}1\}g(X)]$ are estimable by reweighting confirmed claims by $\hat{\rho}$. The post-audit sensitivity of Table 25 operationalizes the same correction at the level of the label rule rather than the point estimate.

Proposition F5 (Trivial-deleter baseline for Edit F1)**.** *Let $n_c, n_s \geq 1$ be the numbers of CVH and supported claims and $\pi = n_c/(n_c + n_s)$ the CVH prevalence, with every editor scored on all $n_c + n_s$ claims and per-claim abstentions counted as deassertions. The delete-all editor attains $\mathrm{EditF1}_\varnothing = \frac{2\pi}{1+\pi}$. Moreover, an editor with CVH deassertion recall r and supported-claim loss q, one minus retention, dominates this floor iff $r > \pi + (1-\pi)\, q$. Editing claims must therefore clear an explicit, prevalence-dependent floor while reporting retention, which is why Sec. 6.2 co-headlines supported-claim retention.*

G Full Proofs

Lemmas G1–G2 below restate the statements of App. C with complete proofs, and Lemma G3 additionally records the point-estimate consequence; the lemma counter restarts in this appendix, and these are restated statements, not new results. Lemmas G4–G8 are the full proofs of the Propositions F1–F5 stated in App. F.

Lemma G1 (Per-occurrence label invariance)**.** *Let a record map each image i to a sequence of spans $s_{i,1}, \ldots, s_{i,m_i}$, each span carrying an authoritative label $\ell_{i,k} \in \{0,1\}$ and a canonical object key $o_{i,k}$. Define the per-(image,canonical-object) population as $\mathcal{P} = \bigsqcup_i \{(i,k)\}$ collapsed by grouping on $(i, o_{i,k})$ with the rule: a group enters the population iff all its members share one label; mixed-label groups are dropped and reported. Then, for any permutation π of the spans within an image, the population computed from $\pi(s_{i,\cdot})$ equals the population computed from $s_{i,\cdot}$.*

Proof. Grouping and labeling are both functions of multiset content: the group key (i, o), the set of labels in a group, and the uniformity test ($\mathrm{card}\{\ell_{i,k}\} = 1$) do not depend on order. The loader's positional FIFO alignment pairs each occurrence with the co-located label, so no label is re-derived through any text-keyed aggregate, and the conflict gate reports, rather than silently keeps, mixed groups. Hence the population is permutation-invariant, and conflicting groups are countable and disclosed. □

Lemma G2 (Calibrated quantile coverage under exchangeability)**.** *Suppose calibration scores $X_1, \ldots, X_n$ are exchangeable, $n \geq 1$, and the threshold is the split-conformal quantile $\hat{\tau} = Q_{\lceil (n+1)(1-\alpha) \rceil / n}(X_{1:n})$ for a target level $\alpha \in (0,1)$, with $\hat{\tau} = +\infty$ in the event $\lceil (n+1)(1-\alpha) \rceil > n$. Let $X_\star$ be a fresh score from the same exchangeable law. Then $\Pr(X_\star \leq \hat{\tau}) \geq 1 - \alpha$, and if the score law is continuous, $\Pr(X_\star \leq \hat{\tau}) < 1 - \alpha + \frac{1}{n+1}$.*

Proof. Exchangeability implies that the rank of $X_\star$ among the joint sample is uniform on $\{1, \ldots, n+1\}$ when scores are tie-free; with ties, the coverage event only contains the rank event, so the lower bound still holds. On the event $\{X_\star \leq X_{(k^\star)}\}$ with $k^\star = \lceil (n+1)(1-\alpha) \rceil$, the probability is $k^\star/(n+1) \geq 1 - \alpha$. The complementary direction follows from the same uniform rank: the atom at the boundary $X_{(k^\star)}$ vanishes under continuity, and only the $1/(n+1)$ discretization term remains, hence the upper inequality is strict. This is the standard split-conformal argument (Lei et al., 2018; Angelopoulos & Bates, 2023); we state the exchangeability assumption explicitly and do not assert it beyond it. □

Lemma G3 (Verdict implication of the break rule)**.** *Let the pre-registered rule be: the nominal guarantee is broken iff the lower bound L of a 95% image-cluster bootstrap CI for $\hat{R}$ satisfies $L > \delta$ (with $\delta = 0.05$). Since $\hat{R} \in [0,1]$ and, in the frozen execution, $\hat{R} = 1$ with every non-empty bootstrap replicate equal to 1 since all audited residual occurrences pass τ, the interval is degenerate with $L = \hat{R} = 1$. Hence whenever the rule fires, the observed point estimate satisfies $\hat{R} > \delta$ too (Efron & Tibshirani, 1993).*

Proof. $L \le 1 = \hat{R}$ because every replicate value lies in $[0,1]$; in the realized case $L = \hat{R} = 1$ degenerately. A firing requires $L > \delta$, so $\hat{R} \ge L > \delta$. The rule therefore implies both the CI-level statement and the point-level statement. □

This is a guard statement about the verdict rule, not a probabilistic calibration claim about $\hat{R}$.

Lemma G4 (Ensemble variance contract; Proposition F1)**.** *Let E_1, E_2 have variances σ_1^2, σ_2^2 and covariance c, and $\bar{E} = (E_1 + E_2)/2$. Then $\mathrm{Var}(\bar{E}) = \frac{1}{4}(\sigma_1^2 + \sigma_2^2 + 2c) \le \frac{1}{2}(\sigma_1^2 + \sigma_2^2)$, with equality iff the two scores are perfectly positively correlated and $\sigma_1 = \sigma_2$.*

Proof. Bilinearity of covariance gives the identity. The inequality is $2c \le \sigma_1^2 + \sigma_2^2$; by Cauchy–Schwarz $c \le \sigma_1\sigma_2$ and by AM–GM $\sigma_1\sigma_2 \le \frac{1}{2}(\sigma_1^2 + \sigma_2^2)$, so $2c \le \sigma_1^2 + \sigma_2^2$. Equality requires both steps tight: perfect correlation $c = \sigma_1\sigma_2$ and $\sigma_1 = \sigma_2$. □

Lemma G5 (Pairing can only improve contrast variance; Proposition F2)**.** *For scores s', s on identical claims, $\mathrm{Var}(s' - s) = \mathrm{Var}(s') + \mathrm{Var}(s) - 2\mathrm{Cov}(s', s)$, which is at most $\mathrm{Var}(s') + \mathrm{Var}(s)$ whenever $\mathrm{Cov}(s', s) \ge 0$.*

Proof. Direct expansion of $\mathrm{Var}(s' - s) = \mathbb{E}[(s' - s)^2] - (\mathbb{E}[s' - s])^2$. The sign condition on the covariance decides the comparison against the sum-of-variances benchmark of independent estimation. Applied to the AUROC contrast of Eq. (7), take s' and s to be the image-level influence aggregates of the two estimators; then $\mathrm{Var}(s' - s)$ is the first-order paired-contrast variance, and a non-negative covariance of the two score aggregates tightens it relative to independent estimation. □

Lemma G6 (Wilson interval form (Wilson, 1927); Proposition F3)**.** *The endpoints of Equation* (19) *lie in $[0,1]$ and increase strictly in k; at $k = 0$ the lower endpoint is 0 and at $k = n$ the upper endpoint is 1 (the documented pins).*

Proof. The Wilson endpoints are the two roots of $n(p - \hat{p})^2 = z^2 p(1-p)$, i.e. $f(p) = (n + z^2)p^2 - (2n\hat{p} + z^2)p + n\hat{p}^2 = 0$. The parabola opens upward with $f(0) = n\hat{p}^2 \ge 0$, $f(1) = n(1-\hat{p})^2 \ge 0$, and axis $(2n\hat{p} + z^2)/(2(n + z^2)) \in [0,1]$, so both roots lie in $[0,1]$. Strict monotonicity in k follows from the implicit derivative $dp/d\hat{p} > 0$ along each root branch of $f(p;\hat{p}) = 0$ on $[0,1]$. At $\hat{p} = 0$ the roots are $\{0,\ z^2/(n+z^2)\}$, giving the lower pin 0; at $\hat{p} = 1$ the roots are $\{n/(n+z^2),\ 1\}$, giving the upper pin 1. Nominal asymptotic coverage follows from the CLT pivot $\sqrt{n}(\hat{p} - p)/\sqrt{p(1-p)} \Rightarrow N(0,1)$. □

Lemma G7 (One-sided audit correction; Proposition F4)**.** *Under one-sided label noise ($z^\star{=}0$ whenever $z{=}0$) with audit-confirmed precision $\rho = \Pr(z^\star{=}1 \mid z{=}1)$ constant over claims, an unbiased audit of the $\{z{=}1\}$ subsample gives $\mathbb{E}[\hat{\rho}\hat{\pi}_z] = \pi_{z^\star}$, where $\hat{\pi}_z$ is the empirical share of claims with $z{=}1$ in the evaluation population, and $\mathbb{E}[1\{z^\star{=}1\}g(X)] = \rho\,\mathbb{E}[1\{z{=}1\}g(X)]$ for every function g of a claim's recorded features X.*

Proof. One-sidedness gives $\Pr(z^\star{=}1, X) = \Pr(z^\star{=}1 \mid z{=}1, X)\Pr(z{=}1, X) = \rho\,\Pr(z{=}1, X)$ by homogeneity of ρ. Marginalizing, $\pi_{z^\star} = \rho\pi_z$. For the estimator, write $N_1 = \sum_i z_i$ and $S = \sum_i z_i^\star$ over the evaluation population. Selection for the audit is unbiased within $\{z{=}1\}$, so $\mathbb{E}[\hat{\rho} \mid \text{population}] = S/N_1$. Conditioning on the population, $\mathbb{E}[\hat{\rho}\hat{\pi}_z] = \mathbb{E}[(N_1/N)(S/N_1)] = \mathbb{E}[S/N] = \pi_{z^\star}$, which requires neither independence of $\hat{\rho}$ and $\hat{\pi}_z$ nor homogeneity of ρ beyond the population identity. The expectation identity follows the same factorization inserted into $\sum_x g(x)\Pr(z^\star{=}1, X{=}x)$. □

Lemma G8 (Trivial-deleter baseline for Edit F1; Proposition F5)**.** *With $n_c \ge 1$ CVH and $n_s \ge 1$ supported claims, every editor scored on all $n_c + n_s$ claims, and per-claim abstentions counted as deassertions, the delete-all editor has $\mathrm{EditF1}_\varnothing = \frac{2n_c}{2n_c + n_s} = \frac{2\pi}{1+\pi}$, where $\pi = n_c/(n_c + n_s)$ is the CVH share among evaluated claims, and an editor with recall r and supported loss q beats it iff $r > \pi + (1-\pi)q$.*

Proof. Delete-all achieves $\mathrm{TP} = n_c$, $\mathrm{FN} = 0$, $\mathrm{FP} = n_s$, so $\mathrm{EditF1} = 2n_c/(2n_c + n_s)$; substitute $\pi = n_c/(n_c + n_s)$. For a general editor with t true positives and f supported losses, dominance is $2t/(t + f + n_c) > 2n_c/(2n_c + n_s)$, which rearranges to $t(n_c + n_s) > f n_c + n_c^2$; dividing by $n_c(n_c + n_s)$ with $r = t/n_c$ and $q = f/n_s$ gives $r > \pi + (1-\pi)q$. □

Table 8: Stage-1 audit: pre-audit CVH instances demoted by the frozen DINO cache at 0.35, the audit-only value. Cache miss = 0 in every split. Metric definitions are collected in App. U.

split	records	pre-audit CVH	demoted texts	demoted occ.	cache miss
dev	300	53	23	29	0
test	1000	188	98	136	0
holdout	2100	298	138	167	0
total	3400	539	259 (48.1%)	332	0

H Stage-1 Detector-Triage Audit

Stage-1 audit re-thresholds the frozen Grounding DINO cache at the audit value 0.35; audit-time demotions quantify incomplete-annotation mislabels. Nothing method-side is used; the complete audit summary is the per-split accounting in Table 8.

The demotion share is our measured estimate, on the audited candidate set, of the annotation incompleteness the audit responds to (Sec. 8); it feeds the audited-variant sensitivity rows (App. L and App. Y) under the audit-independence constraint of Sec. 3.

Of the 259 machine-demoted texts, 157 entered the audit packet and, after removing unresolvable rows, 153 were rated (App. Y). The read-only probe prediction, recorded before the audit ran, which gave 23, 98, and 138 demoted texts on dev, test, and holdout, matches the measured values exactly and serves as the reproducibility checkpoint.

I Backbone-Level Views

Two descriptive views over the five backbone columns. Figure 2 reports the CVH rate (CVH objects per evaluated object claim) per backbone and split. Rates sit in a 0.087–0.195 band across LLaVA, three Qwen3-VL sizes, and gemma: the blind spot appears in every measured backbone column rather than in a single model.

Figure 3 shows the median frozen three-term risk of Eq. (6) for CVH vs supported objects on dev300 across the five columns; the frozen three-term rule is a validation-family context here, and the LLaVA column is the only one with measured model C/U. The gap between the two medians is the signal the risk ranking exploits. Both panels are descriptive: labels are column-native (App. D), so cross-column deltas are not a claim about backbone superiority.

J Audit and Selection Decomposition

Figure 4 decomposes the Stage-1 label audit of Table 8: which pre-audit CVH instances were demoted at the audit threshold, per split. Interpretation follows the audit-independence constraint of the main text and the audit summary of App. H.

Figure 5 decomposes the CEBC selection on eval500, 500 images drawn from the 620-image candidate-generation pool, by generation source — sample, beam, or greedy: candidate captions are produced by sampling, beam search, and greedy decoding and the selector keeps one per image. The guarantee-breach experiment reported in the main text rests on this pool; publishing the source composition keeps the experiment's inputs auditable and is what later allows exact re-selection.

K Benchmark Anchors

Benchmark anchors for the fixed-claim machinery before its own tables: the POPE spread on the primary backbone LLaVA-1.5-7B over its three official settings (Figure 6), the CEBC before/after CHAIR view around the external guarantee stress test (Figure 7), and the out-of-COCO stress leg on Open Images v7 (Figure 8) that checks the pipeline is not COCO-bound. These are context anchors; the registered contrasts remain Table 4–5.

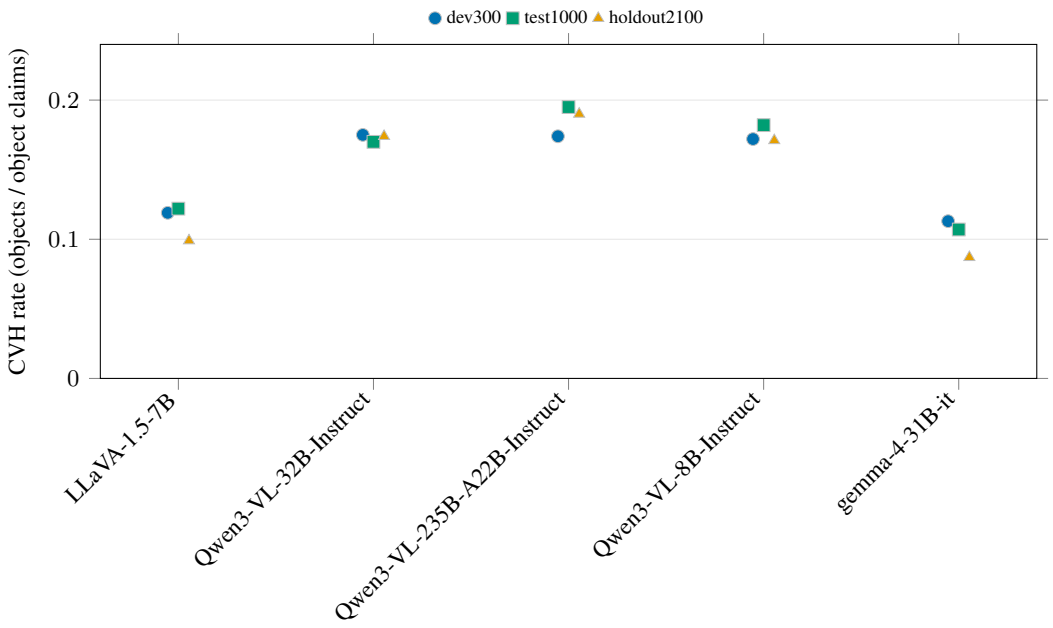

Figure 2: CVH object rate per backbone per split. Rates sit in a tight band (9–20%) across three families and 30× scale; we report this as coverage evidence, not a universality claim. All values are measured under the detector-cache label convention where applicable, and the LLaVA column follows the T1 protocol.

Table 9: Frozen point t=0.73 in bold. Junction-fixed build, regenerated from the frozen pre-junction-fix run artifacts, on the holdout pre-audit population of 3,006 object claims, CVH 298. The flag rule is the frozen ensemble risk $\mathcal{R}_i \geq t$; the frozen 0.73 row reproduces the T2 holdout values. Operating-point sensitivity context only — the registered confirmatory contrast (Table 4) is threshold-free. Metric definitions are collected in App. U.

t	CVH recall	supported specificity
0.50	0.963	0.643
0.60	0.856	0.841
0.65	0.789	0.902
0.70	0.691	0.941
0.73	**0.614**	**0.954**
0.80	0.423	0.979
0.85	0.258	0.992
0.90	0.168	0.995

The detector-row values at the single frozen operating point: detector-cache columns in Table 19, LLaVA column in Table 14. All numbers trace to the frozen per-split builds.

L Sensitivity: Operating Point and Detector Count

Everything in this section varies a knob of the deployed configuration; none of it is a new claim. The registered headline is threshold-free (Table 4); the sweep below (Figure 9) exists to show how the frozen point t=0.73 sits inside the recall/specificity trade-off. The detector-count subsection traces AUROC as detectors are added on the validation-side dev split (Figure 10); the measured one/two/three-detector rows on holdout, which show saturation at two, are Table 12. The risk-family grid is validation-side by decision, and only the dev best point is published (Table 11).

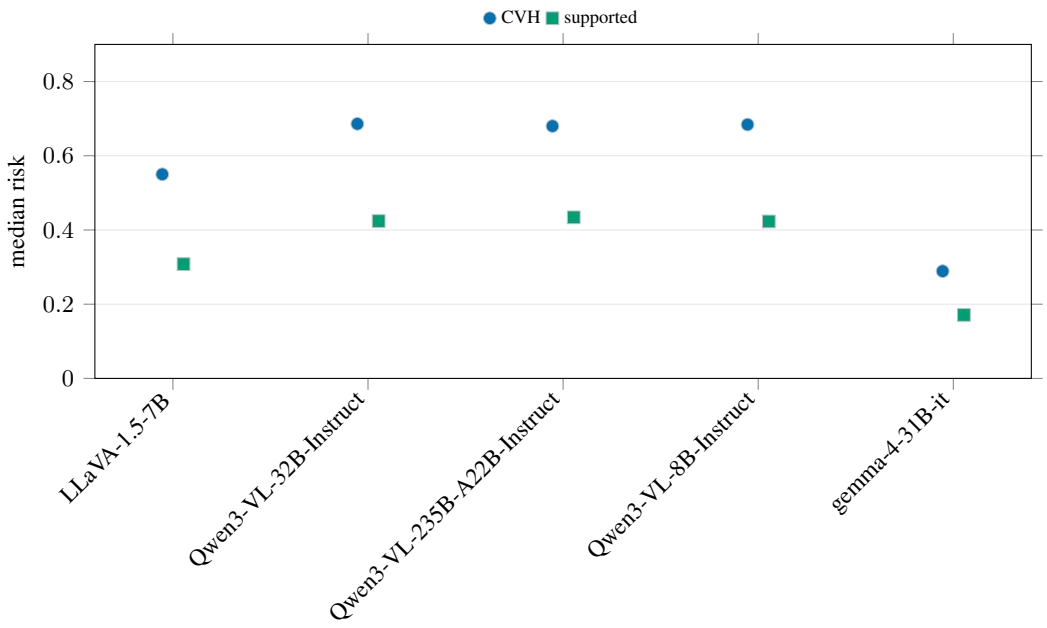

Figure 3: Median frozen three-term risk on CVH vs supported rows on dev300; test1000 and holdout2100 show the same ordering in every column. The CVH band sits above the supported band in all five columns; the gemma column lacks model-measured U/C by convention, so the risk omits the U term—its level shift is a convention difference: gemma's U/C-dependent inputs are n/a on detector-cache columns by design (see the caption of Table 19 and App. W).

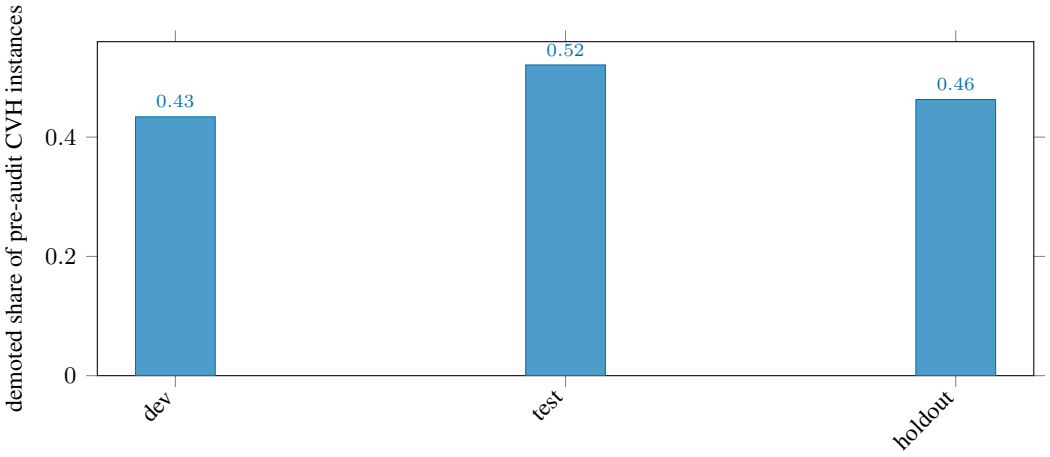

Figure 4: Stage-1 audit demotion share on the texts basis — 23 of 53 on dev, 98 of 188 on test, and 138 of 298 on holdout. Pre-audit CVH sets are measurably contaminated with annotation-incompleteness mislabels ($\approx$ 48% pooled), quantifying the measurement problem.

L.1 Operating-Point Sweep on the Pre-Audit Holdout

L.2 Detector-Count Ablation

L.3 Risk-Family Weight Grid

The pre-junction-fix loader placed its optimum on the overconfidence-gap term, $(0, 0, 1)$ with dev AUROC 0.9195; the label-noise audit thus explains the earlier preference for the gap term, and we keep the registered equal-weight choice, the junction-fixed tie, and the recorded best point of Sec. 4.

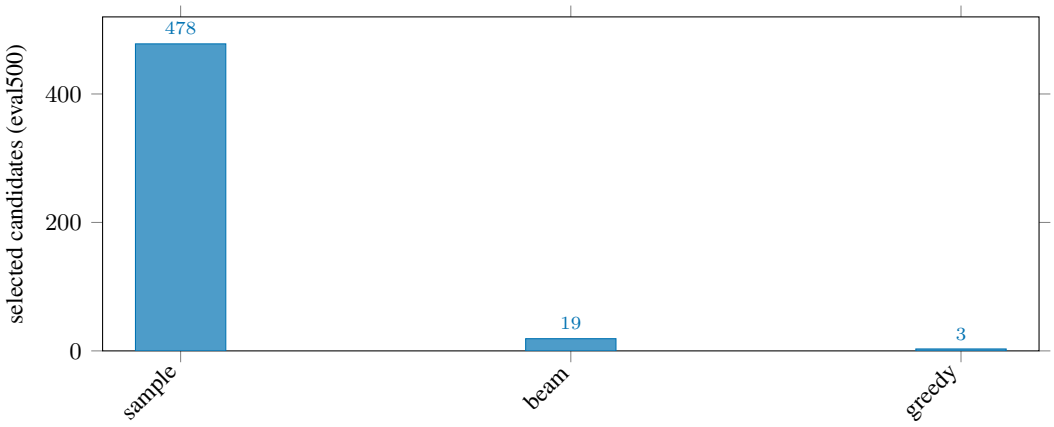

Figure 5: Selection-source composition on eval500, identical on both arms: 478 sample / 19 beam / 3 greedy. Within non-greedy winners, 462 are logp-flip driven; the object reward is not the driver—the measured basis of the fidelity disclosure.

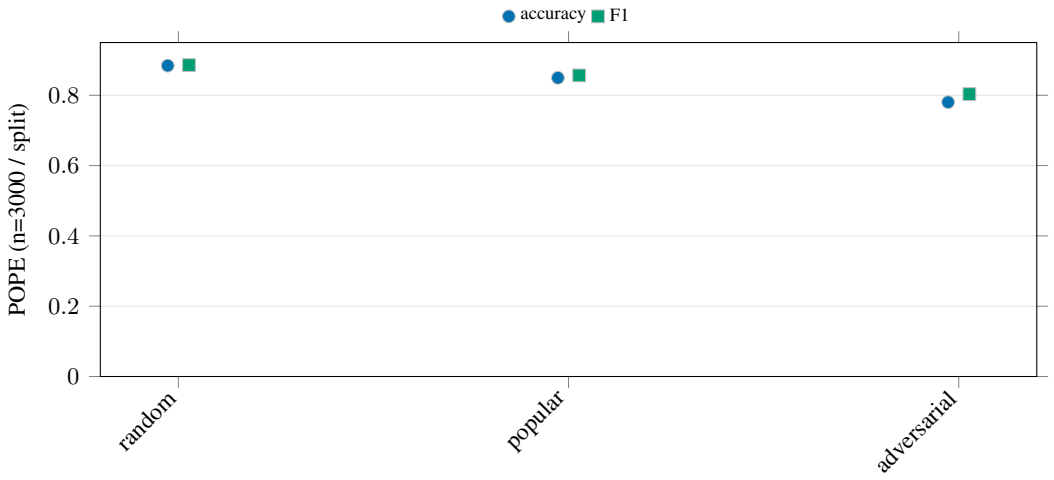

Figure 6: POPE on LLaVA-1.5-7B: accuracy and F1 degrade monotonically random > popular > adversarial, as expected from the POPE design.

Measured cost profile. Per-stage wall clocks are measured from run-event timestamps (Table 16): on the full holdout run the detector stage accounts for 43.3% of wall clock, that is 1,949.6 s of 4,504.4 s, with the dump stage at 56.0%; on the dev300 run the detector stage accounts for 41.7% of the itemized stage total (41.4% of wall clock); detection is roughly 0.93 s per question on holdout. Combining the logged two-detector run with a fresh same-pool OWLv2-only detector run of 1,146 s on the same host class with input identity yields the latency multipliers in Table 12, 0.82 versus 1.00; the second detector family accounts for roughly one-fifth of pipeline wall clock. A third detector checkpoint OWL-ViT v1 adds a measured 1,163 s serial stage, multiplier 1.26, without exceeding the two-detector AUROC, so the deployed configuration stays at two detectors.

M Cause Decomposition

Table 13 decomposes pre-audit high-confidence unsupported spans into a heuristic classification — iscrowd is structurally zero in this build. The "high-evidence" subset collects the spans whose detector evidence $\max(\text{DINO}, \text{OWLv2})$ reaches the audit threshold 0.35. "Residual" cases are accountable to image content alone (the object is simply absent), while "compound" cases involve an interacting factor (e.g. attribute or count error riding on a real object); the class name "residual" denotes a cause

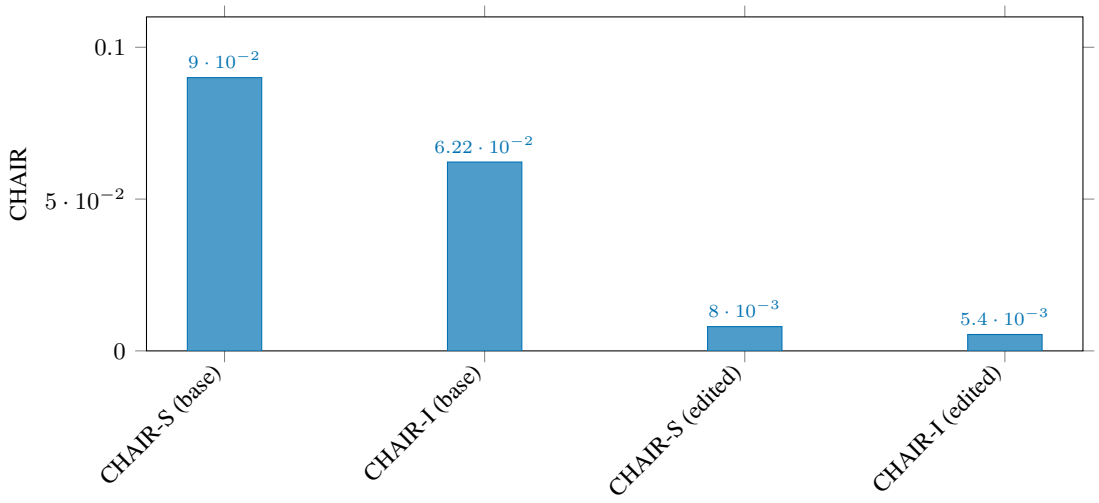

Figure 7: CHAIR before vs after the base/safe-subset edit on eval500 (strict reference gate values used as base). Literal-hallucination metrics collapse to near zero while the distributional guarantee row (R) still breaks—two distinct phenomena.

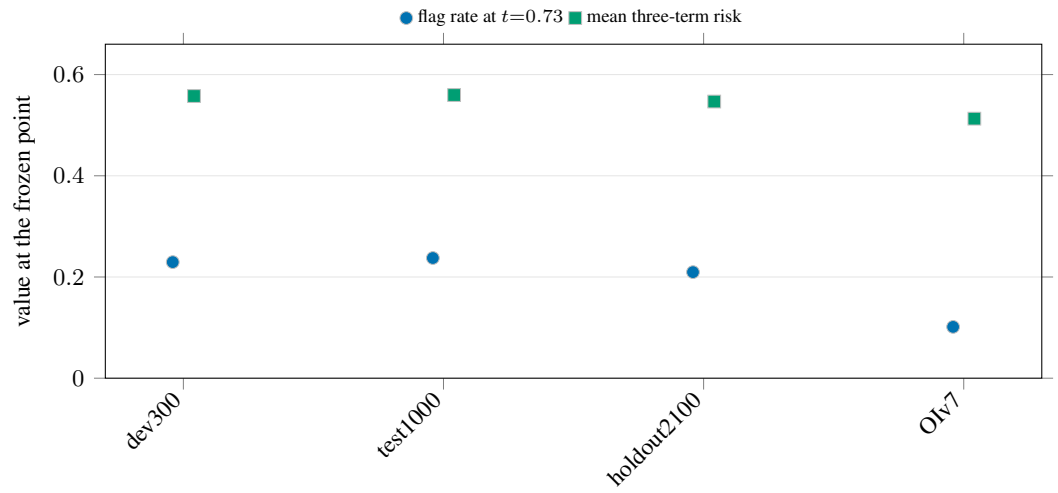

Figure 8: Definition-matched stress comparison at the frozen point: in-COCO splits under the detector-cache convention vs OIv7, the Open Images v7 out-of-COCO leg, grouped by split. The flag rate — the share of claims flagged at t=0.73 — and the mean three-term risk agree in direction; the magnitude of the OIv7 row is the disclosed caveat.

class here and is unrelated to the residual risk $\hat{R}$ of Eq. (17). Residual dominates every split (85–92%), which is what motivates an evidence-first detector over attribute-level correction.

Two reading constraints. The shares are a heuristic decomposition, not a causal claim, and the n column is occurrence-level while the headline CVH counts are object-level (53/188/298); the two grains differ by the collapse map of App. C.

N LLaVA Column: T2 Mechanism

The LLaVA column pre-dates the detector-cache convention and carries measured C/U directly from the original span extraction. Reverse-uncertainty is a frozen-rule row: U-only detection evaluated at its frozen threshold. The *evicvl* row is the legacy bespoke rule kept for lineage; it is excluded

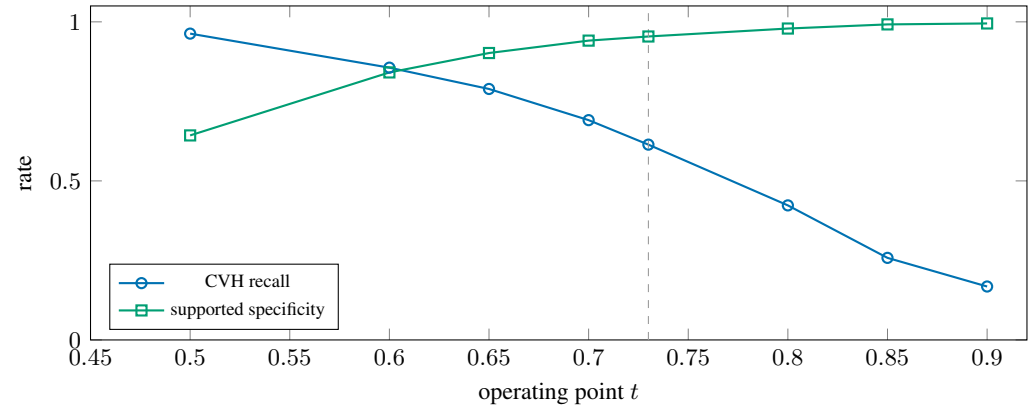

Figure 9: Operating-point sweep on holdout, pre-audit labels (junction-fixed build; values match Table 9). The frozen point t=0.73 is marked; the sweep varies the threshold on the frozen risk of Eq. (5). Sensitivity context only — the registered contrast is threshold-free (Table 4). The x-axis starts at 0.45.



Figure 10: Detector-count contrast on dev300 (junction-fixed lineage; pre-audit n=446 object claims, 53 CVH; post-audit n=423, 30 CVH): the equal-weight ensemble raises the point estimate over the single detector under both label states; the paired intervals are in Table 10. The y-axis starts at 0.80.

from the method claims by the registered lineage decision. All values are from the frozen build. On holdout2100, the reverse_uncertainty, owlv2_only, owlv2_dino_equal, and evicvl rows reproduce Table 5 value for value from the macro-canonical build; only the dino_only holdout row is unique to this table and is shown above.

O Cross-Model Replication Pilot

This section reports the per-column object-level AUROC pilot under column-native labels (Table 15). The pairing grain is the $(\text{image}, \text{canonical object})$ intersection, and the image-clustered bootstrap uses n=1000 replicates with seed 42.

On llava15_7b, the only column with measured model confidence, the uncertainty-direction proxy u_proxy is near chance, at AUROC 0.5277, quantifying the U-only blind spot: objects that the generator claimed confidently sort at the LOW end of an uncertainty-direction score, so an uncertainty-only ranking cannot surface them.

Table 10: 1 vs 2 detectors on dev. ΔAUROC = ensemble − single, image-cluster bootstrap 95% CI (n=3000, seed 0), on the frozen junction-fixed build. Validation-side only; the confirmatory contrast is holdout-only (Table 4). Metric definitions are collected in App. U.

detector count	split	audit	AUROC	ΔAUROC [95% CI]	n object/CVH
1	dev	pre-audit	0.8909	reference	$446/53$
2	dev	pre-audit	0.9238	$+0.0328$ [-0.0004, $+0.0684$]	$446/53$
1	dev	post-audit	0.8720	reference	$423/30$
2	dev	post-audit	0.9626	$+0.0897$ [$+0.0515$, $+0.1345$]	$423/30$

Table 11: Three-term risk-family weights on the dev population: every simplex weight with step 0.05 (231 points) is scored by the AUROC of $\mathcal{R}_{\alpha,\beta,\gamma}$. Junction-fixed lineage. Validation evidence only — no test or holdout claim is attached, and the artifact's holdout field is deliberately unpublished. Metric definitions are collected in App. U.

selection	(α, β, γ)	dev AUROC
grid best	$(0.0,\ 1.0,\ 0.0)$	0.9238

P Additional Quantitative Views

This section presents machine-generated quantitative views over frozen artifacts; every number is re-derived at emission time from the artifact named in each file header. These panels re-visualize values already carried in the main-text tables and add the stage-level cost decomposition behind the T8 latency multipliers.

Panel guide: (i) POPE F1 grouped bars across seven methods plus the gemma backbone-replication column (Figure 11); (ii) CHAIR-S with bootstrap CIs on test and holdout (Figure 12); (iii) T6 per-column detector-risk AUROC with the near-chance u_proxy marker (Figure 13); (iv) the paired ΔAUROC forest with percentile and BCa intervals (Figure 14); (v) the recall–specificity map at the frozen point (Figure 15); (vi) the stage-level wall-clock decomposition as an auditable table (Table 16).

The views below regroup values already carried in the main-text and appendix tables. Additional panels add the all-methods AMBER view (Figure 16), the junction-fixed dev/test sweeps (Figures 18 and 19), the T4 factuality/coverage view (Figure 17), the ConfLVLM guarantee stress view (Li et al., 2025) (Figure 20), and the risk-family weight grid (Figure 21); the per-backbone population view is superseded by Table 7 and is not duplicated here.

Q Protocol Receipts and Provenance Hardening

The confirmatory execution is reproducible from the frozen artifacts alone (the normative identity list is App. R). Pinning of input IDs and content hashes (App. R), the append-only first-fingerprint ledger with drift demoted to exploratory (Sec. 5), and fail-closed caches without zero-imputation (Sec. 4) are stated once at those canonical loci and are not duplicated here. The remaining hardening rule is local to this appendix: every registered evaluation artifact is receipted into a per-day SHA-256 manifest, so the executed state is enumerable and drift between days is detectable. The mezzanine section (App. V) lists the load-bearing receipts for the values in the text.

R Protocol and Artifact Identity

The confirmatory record must contain the following before the first holdout evaluation: role `confirmatory_holdout`; all selected IDs and image hashes; complete past-use denylist hashes; population size and deterministic selection procedure; repository commit; candidate definitions and weights; threshold and configuration hashes; VLM and detector revisions; endpoint, α, bootstrap count and seed; dataset, question, caption, label, cache, and input checksums; and a two-person approval record. Runtime evaluation checks the exact ID and cache-key sets, duplicates, missing

Table 12: T8: detector-count and operating-point sensitivity on the holdout. Candidate selection occurs on validation only. The minimal/full policy variants are omitted rather than marked n/a: the distinguishing flags reach no executable path, so both variants would be behaviorally identical to the frozen policy and a reported row would imply a non-existent distinction. The three-detector row adds OWL-ViT v1, an independent checkpoint and not a third detector family, and does not exceed the two-detector row, the paired increment -0.0018 $[-0.0082, +0.0040]$; detector count saturates at two. Latency multipliers are normalized to the logged two-detector holdout run; stage wall clocks are itemized in the paragraph below. Bold marks the best cell(s) per column; all ties are bold. Registered as T8 in Table 17. Metric definitions are collected in App. U.

Variant	Object AUROC ↑	ΔAUROC ↑	Supported retention ↑	Latency multiplier ↓
One detector	0.887	reference	0.800	**0.82**
Two detectors, equal weight	**0.927**	**+0.0401**	**0.954**	1.00
Three detectors, optional checkpoint	0.925	+0.0382	0.940	1.26

Table 13: Explanation shares per split, all / high-evidence subsets. The n column is occurrence-level; the object-level CVH counts are 53/188/298 by split. Metric definitions are collected in App. U.

split	n CVH	residual	compound	residual, high-evidence	compound
dev300	59	0.915	0.085	0.906 (32)	0.094
test1000	234	0.855	0.145	0.796 (147)	0.204
holdout2100	351	0.866	0.134	0.846 (175)	0.154

values, hashes, revisions, and clean Git state. It never overwrites an existing result. Each identity field is pinned to exactly one value for the entire confirmatory record. The runtime-side receipts and hardening rules are collected in App. Q.

The pipeline figure appears as Figure 1; the other planned pipeline figures are omitted from the printed paper, with App. E retained as a provenance note. The remaining figure-production items are independent of the executed T3 study. We display no synthetic curves or hand-selected examples before their evidence gates pass.

S ConfHall-VL Construction Details

For every image, construction loads the caption trace and COCO annotations, extracts spans, canonicalizes object names with the CHAIR synonym table, and applies Equation (2) per occurrence. For REVERSE-style traces, the spans are the generator's own `<SPAN>` tag phrases, and for plain captions, a frozen lexical scan over the CHAIR synonym table emits single-word mentions and guarded two-word mentions. The frozen record format and the (I, o) collapse with its conflict gate are defined in Sec. 3 and in the per-occurrence label invariance lemma of App. C. The stored editor hash is recomputed before every same-caption comparison, and the implemented loader aggregates each object-level T1 row by maximum confidence and evidence and minimum uncertainty. The released data card reports the exclusion accounting measured from the frozen build inputs: unresolved image IDs, missing caption traces, label conflicts, and the detector-cache audit row are all 0 in every split; the only nonzero exclusion is non-object spans — 302 on dev, 1,131 on test, and 1,775 on holdout. Every build input is receipted with a content checksum, so the exclusion accounting above is recomputable from the frozen inputs alone.

T Human Editing Rubric

This appendix specifies the rating protocol of Sec. 6.2 and reports its executed run over 105 claim occurrences (Table 18): two independent complete passes, raw agreement 0.571, Cohen's κ 0.388 (5-way) and 0.949 (deasserted vs. kept), 138 adjudicated disagreements and 0 unresolved. The on-packet label set {keep, hedge, removed, abstain, unclear} compresses the rubric's negated/omitted into removed; the compression is disclosed here per the registered protocol. Rater use of the unclear label, for items a rater could not resolve, was 29/420 and 13/420 of answers (raters A and B, respectively)

Table 14: T2 mechanism, LLaVA column. Metric definitions are collected in App. U.

split	method	CVH recall	supported specificity
dev300	reverse_uncertainty	0.000	1.000
	owlv2_only	0.830	0.786
	dino_only	0.283	0.975
	owlv2_dino_equal	0.698	0.934
	evicvl	0.717	0.885
test1000	reverse_uncertainty	0.000	1.000
	owlv2_only	0.782	0.773
	dino_only	0.165	0.980
	owlv2_dino_equal	0.569	0.948
	evicvl	0.660	0.876
holdout2100	dino_only	0.268	0.985

Table 15: T6 dev300 pilot, per-column detection under column-native labels. Detector-cache columns derive labels from the frozen ensemble risk, so within-family paired deltas are zero by construction on shared conventions: the share of shared keys whose column labels disagree is exactly zero for every within-family pair. Only cross-convention pairs bear information. The registered cross-convention pair is Table 23. Metric definitions are collected in App. U.

column	images	object claims	CVH	detector-risk AUROC
llava15_7b	262	446	53	0.8909
qwen3vl32b	300	968	169	0.9373
qwen3vl235b	300	1039	181	0.9324
qwen3vl8b	300	1001	172	0.9412
gemma4_31b	297	790	89	0.9514

against the pre-specified 5% guideline: A exceeded it at 6.9% while B stayed within at 3.1%; this exceedance is recorded here per protocol. Of the 138 adjudicated disagreements, 72 were resolved mechanically and 66 by the author rater re-inspecting the paired captions; that the author rater served as adjudicator for these cases is disclosed here per the registered protocol.

For each object-claim occurrence, raters answer: (1) does the base caption positively assert the claim; (2) is the claim visually supported; (3) does the edited caption positively assert, hedge, negate, omit, or abstain on it; and (4) did the edit introduce any new unsupported object claim? Method names, detector scores, CVH/support roles, and other raters' labels are hidden. Two independent complete packets are required. A separate adjudicator resolves all disagreements while preserving both original ratings. Cohen's κ, raw agreement, adjudication count, and unresolved count accompany the headline.

U Metric Definitions

All metrics are computed at the frozen ensemble operating point $t{=}0.73$ (unless stated otherwise) and are evaluated either per span occurrence or per canonical object as denoted. Notation follows Table 6: $z_{io} \in \{0,1\}$ is the CVH label of canonical object o on image i. The primary contrast $\widehat{\Delta}_{\mathrm{AUROC}}$ (Eq. (7)) and the editing metric (Eq. (8)) are defined in Sec. 6; every remaining reported metric is defined below.

Risk score and flag rule. For a detector ensemble of size K (deployed $K{=}2$) with per-backend evidence scores $E_b(o,i) \in [0,1]$, the risk and the flag at operating point t are

$$\mathcal{R}_{i,o}^{(K)} \;=\; 1 - \frac{1}{K}\sum_{b=1}^{K} E_b(o,i), \qquad f_{i,o}(t) \;=\; \mathbb{1}\Big[\mathcal{R}_{i,o}^{(K)} \geq t\Big]. \tag{9}$$

Here $E_b(o,i)$ is the cached detector evidence of Eq. (3) (best post-processed box score for the canonical query, $\max \varnothing{=}0$). Claims without checked evidence are never flagged; they are excluded from the detection population and the exclusion is counted per split.

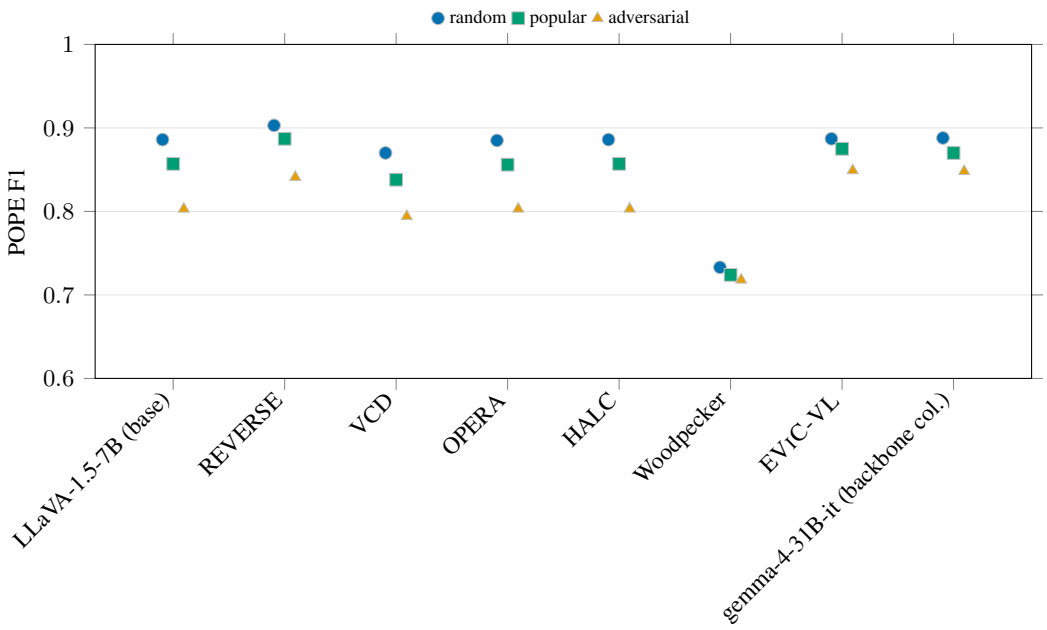

Figure 11: POPE F1 by method across the three official splits, n=3000 per split. gemma-4-31B-it is the T6 backbone-replication column, not an editing method. The y-axis starts at 0.60.

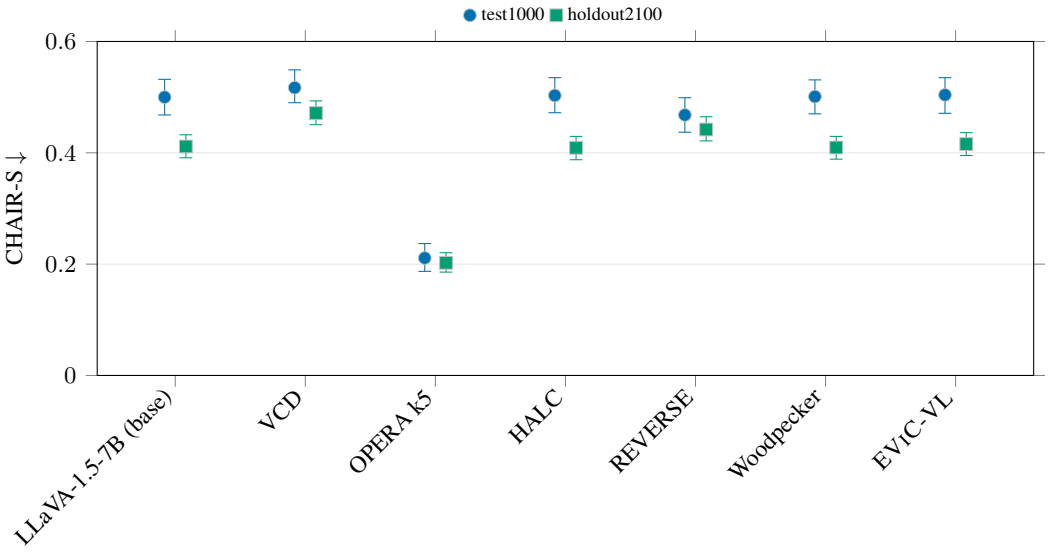

Figure 12: CHAIR-S, the sentence-level hallucination rate where lower is better, with bootstrap 95% CIs (n=1000, seed 42). OPERA is the official-strict k=5 run, matching the OPERA row of Table 21: the figure's test series and the table carry the same OPERA numbers.

CVH-detection recall and supported specificity. Over the object-level evaluation population,

$$\begin{aligned} \mathrm{Rec}(t) &= \frac{\sum_{i,o} z_{io}\, f_{i,o}(t)}{\sum_{i,o} z_{io}}, \\ \mathrm{Spec}(t) &= \frac{\sum_{i,o} (1 - z_{io})\,(1 - f_{i,o}(t))}{\sum_{i,o} (1 - z_{io})}. \end{aligned} \tag{10}$$

"Supported retention" in the tables is $\mathrm{Spec}(t)$ at the frozen point.

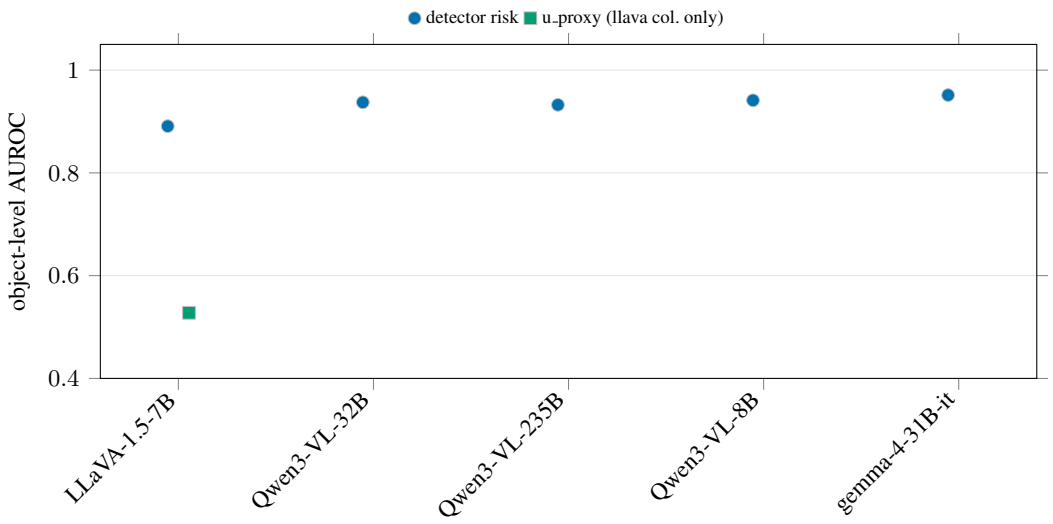

Figure 13: T6 dev300 pilot: per-column detector-risk AUROC under column-native labels. The u_proxy score, the generator's own uncertainty direction, is defined only on the llava15_7b column and sits near chance, quantifying the uncertainty-blindness motif across backbones. The y-axis starts at 0.40.

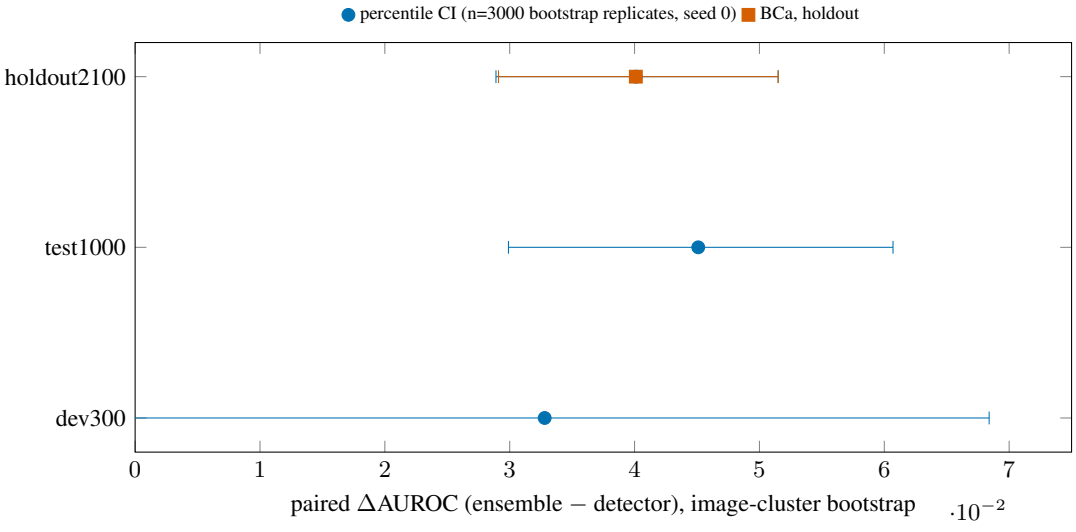

Figure 14: Paired one- vs two-detector ΔAUROC per split with image-cluster bootstrap 95% CIs; deviation from the frozen operating point is excluded by construction. The BCa recomputation overlays the percentile interval on holdout.

Detection metrics and the paired contrast. We evaluate object-level detection of CVH labels: per-occurrence authoritative labels are collapsed to (image, object) pairs to form the evaluation population. Each method's score s ranks claims by that method's own value; with $\mathcal{P}, \mathcal{N}$ the positive and negative object sets,

$$\mathrm{AUROC}(s) = \frac{\sum_{p\in\mathcal{P}} \sum_{q\in\mathcal{N}} \mathbb{1}[s_p > s_q] + \frac{1}{2}\mathbb{1}[s_p = s_q]}{|\mathcal{P}|\,|\mathcal{N}|}. \tag{11}$$

AUPRC summarizes the same ordering in precision–recall space by the non-interpolated average precision (the `average_precision_score` convention). With $\theta_1 > \cdots > \theta_m$ the m distinct

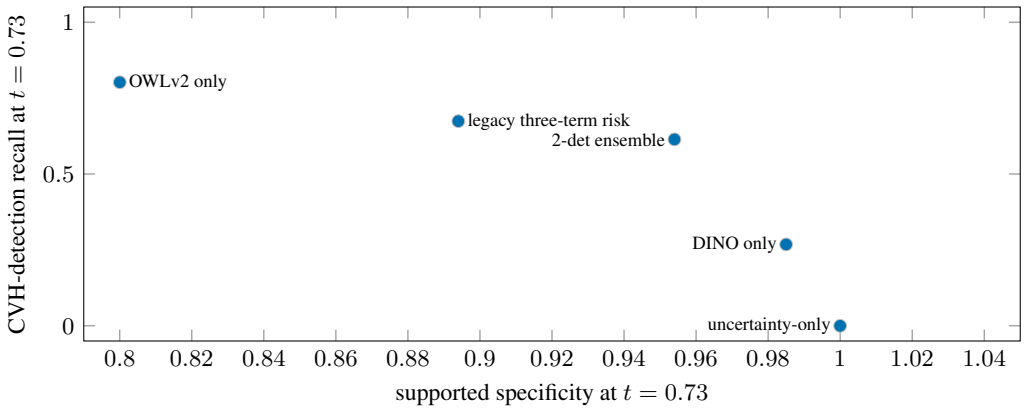

Figure 15: Detection recall vs. supported specificity at the single frozen operating point t=0.73 (holdout2100). The two-detector ensemble trades a moderate recall level for high supported retention; the uncertainty-only rule cannot fire. The "legacy three-term risk" point is the legacy validation family, excluded from method claims (Table 5). The x-axis starts at 0.79.

Table 16: Stage-level wall-clock decomposition of the logged two-detector runs (event-timestamp derived; stages under 1 s omitted, so shares do not sum to 100%). The detector stage share (43.3% holdout) is the measured basis of the T8 latency multipliers in Table 12.

Stage	holdout s	share	dev300 s	share
preflight	13.4	0.3%	10.2	1.3%
dump	2518.3	56.0%	418.0	54.7%
build	16.1	0.4%	16.4	2.1%
detector	1949.6	43.3%	318.4	41.7%
detector-merge	1.6	0.0%	0.8	0.1%

values taken by s on the population and $\mathrm{rec}_0 = 0$,

$$\begin{aligned} \mathrm{AUPRC}(s) &= \sum_{k=1}^{m} (\mathrm{rec}_k - \mathrm{rec}_{k-1})\, \mathrm{prec}_k, \\ \mathrm{rec}_k &= \frac{|\{p \in \mathcal{P} : s_p \geq \theta_k\}|}{|\mathcal{P}|}, \\ \mathrm{prec}_k &= \frac{|\{p \in \mathcal{P} : s_p \geq \theta_k\}|}{|\{o \in \mathcal{P} \cup \mathcal{N} : s_o \geq \theta_k\}|}, \end{aligned} \tag{12}$$

i.e. the weighted mean of per-threshold precisions with recall gains as weights (constant step interpolation). Evaluation counts only at distinct score values, making the value invariant to tie order; if all scores tie, AUPRC equals the positive base rate. Degenerate guards follow the implementation: AUROC is taken as $1/2$ when $\mathcal{P}$ or $\mathcal{N}$ is empty and AUPRC as 0 when $\mathcal{P}$ is empty; the frozen T1 population is non-degenerate. The contrast is paired and per-split, $\widehat{\Delta}_{\mathrm{AUROC}} = \mathrm{AUROC}(s^{(2)}) - \mathrm{AUROC}(s^{(1)})$ with $s^{(2)}$ the equal-weight ensemble and $s^{(1)}$ the single-detector baseline, with an image-cluster bootstrap interval (Cameron & Miller, 2015; Efron & Tibshirani, 1993): images are resampled with replacement B=3000 times (seed 0) and the 95% interval is read from the percentile; the BCa recomputation (Efron, 1987) is reported where stated.

Paired bootstrap intervals. Let $\Delta^*_{(0)} \leq \cdots \leq \Delta^*_{(B'-1)}$ be the order statistics of the paired replicates of Eq. (7): each replicate resamples images with replacement and keeps all object rows of each sampled image, takes the AUROC difference of the two methods on the same resampled rows, and is discarded if the resampled labels are single-class (B=3000 draws, seed 0; B' valid replicates). The reported percentile interval reads the sorted replicates at 0-based floor indices,

$$\mathrm{CI}^{\mathrm{pct}}_{0.95} = \left[\, \Delta^*_{(\lfloor 0.025\, B' \rfloor)},\ \Delta^*_{(\lfloor 0.975\, B' \rfloor)} \,\right], \tag{13}$$

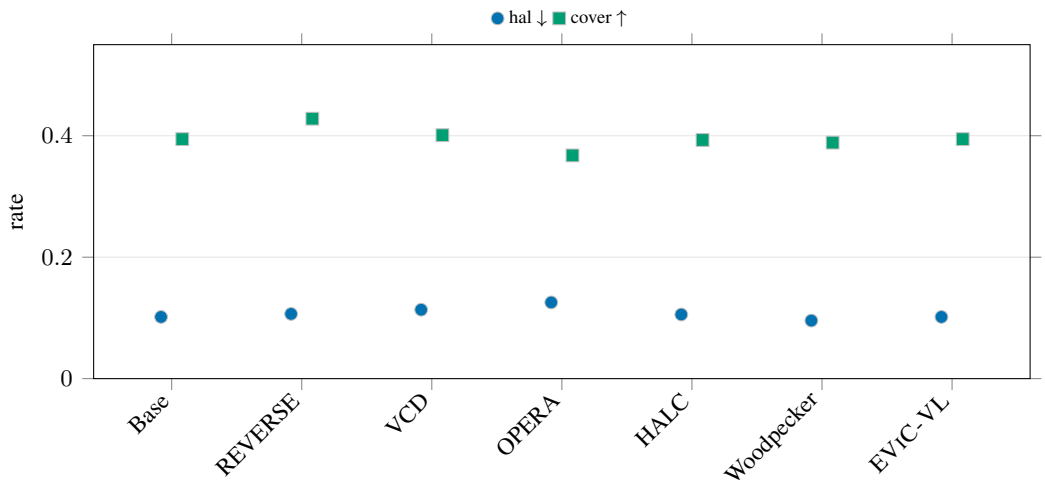

Figure 16: AMBER generative split (n=1004): caption-level hal, where lower is better, and coverage, per method. The EVIC-VL column is the measured run, coinciding with the declared mirror at four decimal places.

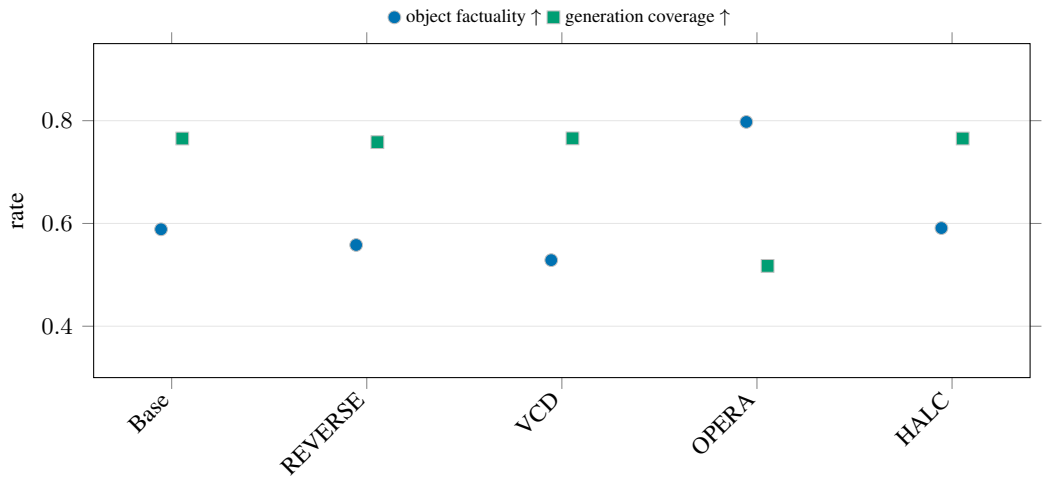

Figure 17: T4 independent-generation factuality and coverage on holdout (n=2100), under the AMBER hal/cover convention transferred to the COCO-GT setting (provisional convention, see Table 21 caption). The y-axis starts at 0.30.

and the reported point estimate is the replicate median, not the plug-in $\widehat{\Delta}_{\mathrm{AUROC}}$. The BCa interval (Efron, 1987; Efron & Tibshirani, 1993) uses the linearly interpolated empirical quantile Q^{lin} with

$$\begin{aligned} \mathrm{CI}^{\mathrm{BCa}}_{0.95} &= \big[\, Q^{\mathrm{lin}}(\alpha_l),\ Q^{\mathrm{lin}}(\alpha_u) \,\big], \\ \alpha_g &= \Phi\!\left(\hat{z}_0 + \frac{\hat{z}_0 + z^{(g)}}{1 - \hat{a}\,(\hat{z}_0 + z^{(g)})} \right), \quad g \in \{l, u\}, \end{aligned} \tag{14}$$

where $\hat{z}_0 = \Phi^{-1}\big(\#\{b : \Delta^*_b < \widehat{\Delta}_{\mathrm{AUROC}}\}/B'\big)$ and $\hat{a} = \sum_i(\bar{\Delta}_{(\cdot)} - \Delta_{(i)})^3 \big/ \big\{6\big[\sum_i(\bar{\Delta}_{(\cdot)} - \Delta_{(i)})^2\big]^{3/2}\big\}$ come from the leave-one-image-cluster-out jackknife $\{\Delta_{(i)}\}$.

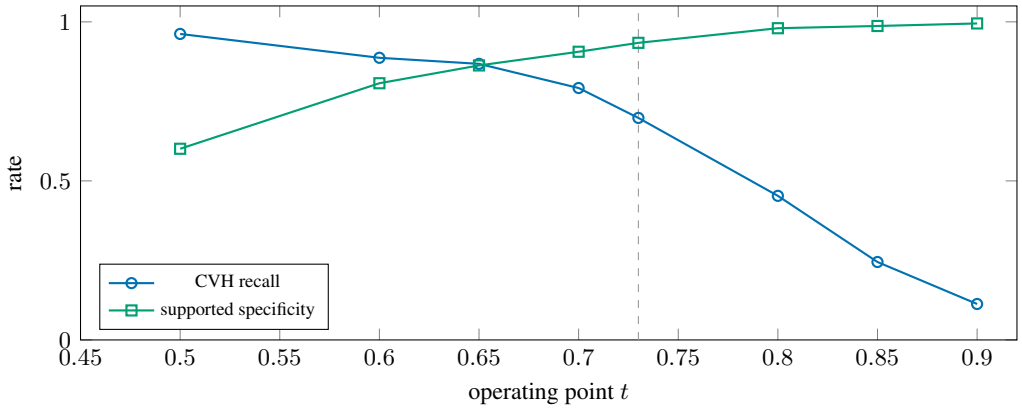

Figure 18: Operating-point sweep on dev300 (junction-fixed build). The frozen point t=0.73 is marked; the sweep varies the threshold on the frozen risk of Eq. (5). Sensitivity context only — the registered contrast is threshold-free (Table 4). The x-axis starts at 0.45.

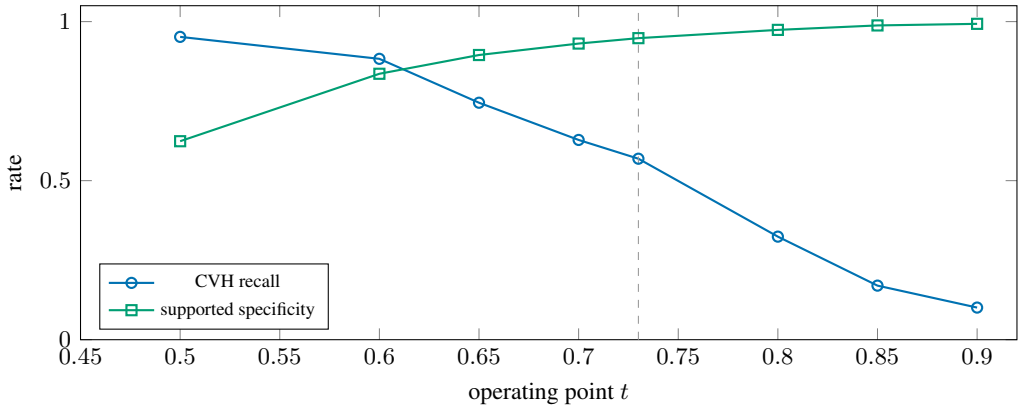

Figure 19: Operating-point sweep on test1000 (junction-fixed build). The frozen point t=0.73 is marked; the sweep varies the threshold on the frozen risk of Eq. (5). Sensitivity context only — the registered contrast is threshold-free (Table 4). The x-axis starts at 0.45.

CHAIR. Per (Rohrbach et al., 2018), with $H(y)$ the number of hallucinated object mentions and $M(y)$ the total mentioned objects in caption y,

$$\text{CHAIR-s} = \frac{\#\{y : H(y) \geq 1\}}{\#\{y\}}, \qquad \text{CHAIR-i} = \frac{\sum_y H(y)}{\sum_y M(y)}; \tag{15}$$

recall is the fraction (micro-pooled over captions) of ground-truth objects mentioned.

POPE. Frozen question–answer pairs per split — random, popular, and adversarial — with the official question formulation; per split we report accuracy and $\text{F1} = 2\,\text{TP}/(2\,\text{TP} + \text{FP} + \text{FN})$ with a bootstrap 95% interval (n=1000, seed 42).

AMBER. For the generative split (n=1004), with c_i the caption for image i,

$$\text{hal} = \frac{\#\{i : H(c_i) \geq 1\}}{1004}, \tag{16}$$

i.e. caption-level CHAIR; the object-level rate is retained in the caption of Table 20. Cover is caption-side ground-truth (GT) coverage and cog is the same measure over the hallucination-prompting side.

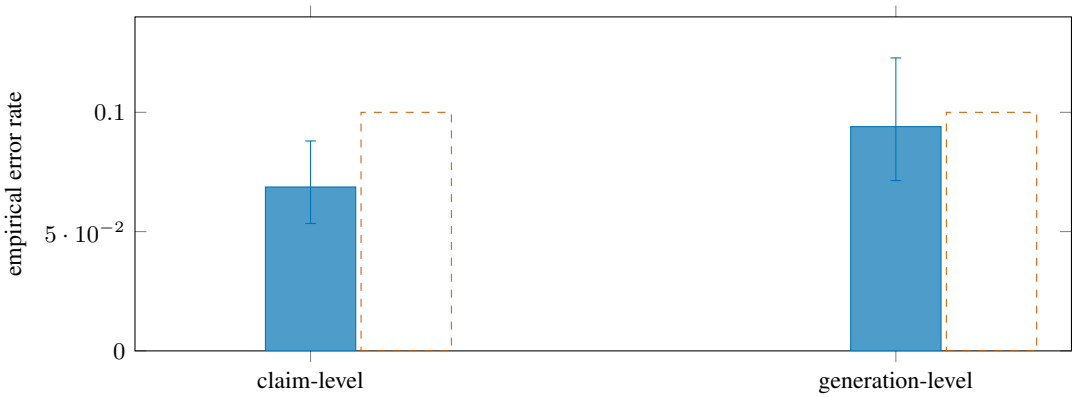

Figure 20: ConfLVLM (Li et al., 2025) re-implementation on the frozen calib120/eval500 split, the 120-calibration/500-evaluation design of Eq. (17), at nominal coverage $1-\alpha$=0.9: the calibrated threshold saturates ($\tau=-\infty$, no claim dropped). The post-filter point estimates sit below α (claim-level 0.069 [0.053, 0.088], generation-level 0.094), but the generation-level 95% CI upper end crosses α=0.100, so the nominal guarantee is not confirmed at the generation level here. The CEBC guarantee was measured broken in the main text — guarantee claims are system-specific, which is why we report both.

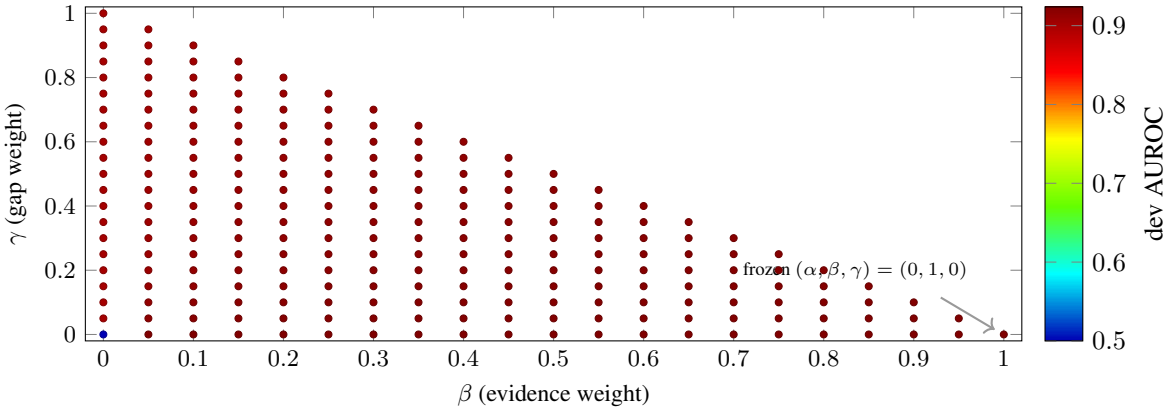

Figure 21: Risk-family weight grid over the simplex $\alpha+\beta+\gamma$=1 (step 0.05, 231 points), junction-fixed dev lineage. Every point on the evidence-only edge (γ=0, β>0) ties at the dev optimum (0.9238); the recorded best point is the frozen equal-weight choice (0,1,0). Validation-side evidence only — no holdout field is published for this grid.

Detection residuals and revision rate. With τ the calibrated evidence threshold, let J be the set of mention occurrences on the final post-edit captions that are absent from the evaluation ground truth, and E_j the detector evidence score of mention j,

$$\hat{R} = \frac{\#\{j \in J : E_j \geq \tau\}}{\#J}, \qquad \text{Rev.} = \frac{\#\{i : y_i^* \neq y_i\}}{|I|}, \tag{17}$$

i.e. $\hat{R}$ is the share of ground-truth-absent post-edit mentions that still pass the evidence test at τ — the residual hallucinated mass the calibrated gate failed to remove — and Rev. is the fraction of images whose final caption differs from the greedy base caption (exact string inequality after whitespace normalization).

Calibration of τ. For the external CEBC stress test, τ is the pre-registered $(1-\delta)$-quantile of the calibration-split hallucinated-mention scores, $\tau = \min\big(\max(Q_{1-\delta}(H),\, 0.25),\, 0.95\big)$ with δ=0.05, fallback τ=0.80, and calibration size n_{cal}=120; here $Q_{1-\delta}$ is the type-7 linearly interpolated quantile, whereas the guard lemma (Lemma G2) attaches its coverage statement to the $\lceil (n_{\text{cal}}+1)(1-\delta) \rceil$-th

Table 17: Paper-facing experiment contract. "Primary" denotes the sole prospective confirmatory detector comparison, not the importance of other scientific questions. T1/T2/T4/T5/T6/T8 cells are populated from frozen artifacts, and the T7 cells from the executed blinded audit. The T3 cells come from the executed 105-item blinded study (two raters, adjudicated; App. T). Result tables by track: T1 → Table 4, T2 → Table 5, T3 → Table 18, T4 → Table 21, T5 → Table 22, T6 → Table 23, T7 → Table 24, T8 → Table 12. Metric definitions are collected in App. U.

ID	Question	Population / methods	Headline	Validity gate
T1	Prospective detection	2,100 images; OWLv2 vs. equal ensemble	paired ΔAUROC	frozen fingerprint + ledger
T2	Fixed-claim blind spot	U-only, OWLv2, ensemble, legacy three-term rule	CVH recall	identical claim identities
T3	Same-caption editing	detector revision, EVIC-VL	Edit F1	105-item blinded study, 2 raters + adjudication
T4	Independent generation	base, REVERSE, VCD, OPERA, HALC	factuality / coverage	no edit ranking
T5	Standard benchmarks	CHAIR, POPE, AMBER	benchmark metrics	official revisions, complete outputs
T6	Cross-model replication	at least two frozen VLM revisions	ΔAUROC	same frozen protocol
T7	Label quality	stratified blind audit	precision + Wilson score CI (Wilson, 1927)	blinded raters + control probes + washout re-rate
T8	Sensitivity and cost	detector count, threshold, components	AUROC / retention / latency	validation only

Table 18: Blinded human editing study (T3) on 105 same-caption item pairs. Edit F1 follows Eq. (8) with deassertion = hedge/removed/abstain; control damage counts interventions on supported (control) claims (denominator 24). Two independent raters, adjudicated; 5-way $\kappa = 0.388$, binary $\kappa = 0.949$, raw agreement 0.571. Edited captions are the frozen bundle generation. Registered as T3 in Table 17.

Method	Edit F1	Precision	Recall	Control damage
detector revision	0.7500	0.7857	0.7174	9/24
EVIC-VL	0.7792	0.9677	0.6522	1/24

order statistic, $k^\star$=115 at n_{cal}=120, one rank above the deployed value — the convention difference is disclosed rather than hidden. Both dual-GT calibration arms P and C return the identical value τ=0.7197; neither the fallback nor the clipping bound was triggered. The statistical-power disclosure is collected in App. V.

Editing metrics. The editing metric and its secondary descriptors are computed over claim-occurrence rating records (rubric: App. T). With TP, FN, FP as defined in Sec. 6.2 and TN the supported claims still positively asserted after the edit,

$$\begin{aligned} \text{EditF1} &= \frac{2\text{TP}}{2\text{TP}+\text{FP}+\text{FN}} \text{ (Sec. 6.2)}, \qquad r = \frac{\text{TP}}{\text{TP}+\text{FN}}, \\ \text{retention} &= \frac{\text{TN}}{\text{TN}+\text{FP}}, \qquad \text{strict} = \frac{N_{\text{neg}}+N_{\text{omit}}}{\text{TP}+\text{FN}}, \end{aligned} \tag{18}$$

Retention here is distinct from the supported retention $\text{Spec}(t)$ above. Here r is CVH deassertion recall and hedging counts toward the r numerator but never toward strict, whose numerator instead counts CVH claims negated or omitted (a grounded replacement is observed as an omission of the original claim). The remaining descriptors are: the new unsupported object-claim rate, the share of evaluated captions whose edit introduces at least one new unsupported object claim (rubric question 4); evaluation coverage, the share of claims not abstained with abstentions counted as kept claims in the TP/FN/FP/TN accounting, hence an all-abstain editor is scored rather than excluded,

#{claims not abstained}/(TP+FN+FP+TN); and the abstention rate, the share of evaluated captions fully abstained.

Proportion intervals for the label audit. Binomial proportions use the Wilson score interval (Wilson, 1927) at level $1-\alpha$:

$$\frac{\hat{p} + \frac{z^2}{2n} \pm z\sqrt{\frac{\hat{p}(1-\hat{p})}{n} + \frac{z^2}{4n^2}}}{1 + \frac{z^2}{n}}, \qquad z = \Phi^{-1}(1-\alpha/2), \tag{19}$$

with the documented exact edge pins low=0 at k=0 and high=1 at k=n.

Inter-rater agreement. With p_o the observed agreement and p_e the chance agreement over the rubric's marginals, Cohen's $\kappa = (p_o - p_e)/(1 - p_e)$. For the executed audit, κ is computed on the 160-row shared subset rated by the second rater, restricted to the N=152 rows where both judgments are present or absent.

Label-audit rates. Audit judgments are $\mathrm{judg}_r(x) \in \{\text{present}, \text{absent}, \text{unresolvable}\}$; rate denominators exclude unresolvable items unless stated. For a stratum $\mathcal{S}$ and rater A (the author),

$$\mathrm{absent}_{\mathcal{S}} \;=\; \frac{\#\{x \in \mathcal{S} : \mathrm{judg}_A(x) = \text{absent}\}}{\#\{x \in \mathcal{S} : \mathrm{judg}_A(x) \in \{\text{present}, \text{absent}\}\}}, \tag{20}$$

so that the headline CVH precision is $\mathrm{absent}_{\mathrm{retained}}$=180/196, the demoted candidates confirmed present are $1-\mathrm{absent}_{\mathrm{demoted}}$=43/153, and the pooled confirmed-absent share is (180+110)/(196+153)=290/349; all proportion intervals use Eq. (19). The control-probe false-alarm rate is $\mathrm{absent}_{\mathrm{control}}$=80/439; controls never enter the precision denominator. On the shared subset, the second rater's corresponding absent rate is 35/57, with the denominator counting only the second rater's present/absent judgments. Self-agreement is the exact-identity repeat rate on the author's 20% washout re-rating, 147/159, counting unresolvable judgments as judgments, i.e., with no exclusions. All audit provenance and population details are in App. Y.

Latency multiplier. With W_{nd} the non-detector wall clock and w_b a fresh, same-pool, serial stage wall for backend b on the same host class,

$$\ell_K = \frac{W_{\mathrm{nd}} + \sum_{b=1}^{K} w_b}{W_2}, \qquad \ell_2 \equiv 1, \tag{21}$$

normalized to the logged two-detector holdout run ($W_2 = 4{,}504.4$ s; $w_{\mathrm{OWL}} = 1{,}146$ s, $w_{\mathrm{OWLv1}} = 1{,}163$ s).

Mechanism rows. CVH-detection recall and supported specificity follow Eq. (10) at the frozen point per rule; U/C-dependent rows are n/a unless U and C are restored from the measured token stream — case A, with the naming in App. W. The median-risk columns are context computed once on the identical fixed-claim population (3,006 object claims; 298 CVH, 2,708 supported): each claim (i, o) is scored by the three-term validation family of Eq. (6) at the frozen weights (α, β, γ)=$(0.25, 0.35, 0.40)$ with E=E_{ens} (Eq. (4)), clipped to $[0, 1]$, and the columns report the label-stratified medians

$$\underset{(i,o):\, z_{io}=1}{\mathrm{median}} \; R_{i,o}^{\text{three-term}} \qquad \text{and} \qquad \underset{(i,o):\, z_{io}=0}{\mathrm{median}} \; R_{i,o}^{\text{three-term}}. \tag{22}$$

No operating point is applied, so every score row of the table shares the same two values by construction. They summarize the distribution of the shared validation risk as context and attribute no result to the risk family itself.

Label audit. The demotion rate is the share of pre-audit CVH labels demoted, the occurrence collapse is the share of duplicate-text occurrences that collapse to one object, and category-structure shares are computed over the audited populations; κ is reported when two independent raters are present.

Table 19: Mechanism analysis at the frozen operating point t=0.73: CVH-detection recall / supported specificity per evidence rule and backbone. A C/U-dependent row is n/a on a detector-cache column unless restored: the case A blocks restore both signals from the saved token stream with 100% join coverage, so their last two value groups carry reverse-U and three-term-risk rec./spec., not owlv2/dino; the case C dev300 row uses the post-fix span extraction and loses one joined claim: 967/168 vs Table 7's 968/169. Case naming and the restore construction are in App. W. The informative cells are the single-detector complementarity. Metric definitions are in App. U.

backbone	split	n_{objects}	CVH	owlv2 rec./spec.	dino rec./spec.
Qwen3-VL-32B-Instruct	dev300	968	169	0.982/0.731	0.408/0.974
Qwen3-VL-32B-Instruct	test1000	3147	534	0.985/0.728	0.335/0.976
Qwen3-VL-32B-Instruct	holdout2100	6625	1152	0.988/0.732	0.330/0.978
Qwen3-VL-235B-A22B-Instruct	dev300	1039	181	0.983/0.728	0.403/0.966
Qwen3-VL-235B-A22B-Instruct	test1000	3379	659	0.989/0.723	0.323/0.975
Qwen3-VL-235B-A22B-Instruct	holdout2100	7052	1342	0.987/0.722	0.354/0.976
Qwen3-VL-8B-Instruct	dev300	1001	172	0.988/0.728	0.384/0.972
Qwen3-VL-8B-Instruct	test1000	3231	587	0.991/0.723	0.308/0.977
Qwen3-VL-8B-Instruct	holdout2100	6679	1142	0.988/0.729	0.355/0.980
gemma-4-31B-it	dev300	790	89	0.989/0.759	0.382/0.983
gemma-4-31B-it	test1000	2590	277	0.996/0.746	0.260/0.984
gemma-4-31B-it	holdout2100	5136	447	0.991/0.762	0.266/0.985
case A rows, U/C restored: the last two value groups are reverse-U and three-term-risk rec./spec., not owlv2/dino:					
Qwen3-VL-8B-Instruct	dev300	1001	172	0.802/0.321	0.343/0.993
Qwen3-VL-8B-Instruct	test1000	3231	587	0.828/0.298	0.264/0.992
Qwen3-VL-8B-Instruct	holdout2100	6679	1142	0.808/0.323	0.288/0.992
Qwen3-VL-235B-A22B-Instruct	dev300	1039	181	0.845/0.262	0.315/0.990
Qwen3-VL-235B-A22B-Instruct	test1000	3379	659	0.818/0.299	0.287/0.993
Qwen3-VL-235B-A22B-Instruct	holdout2100	7052	1342	0.800/0.343	0.288/0.990
Qwen3-VL-32B-Instruct, case C	dev300	967	168	0.833/0.302	0.363/0.992

Generation coverage and object factuality. Scored on the frozen T4 dumps under the AMBER hal/cover convention transferred to the COCO ground truth: hal is the fraction of generations with at least one GT-inconsistent object mention, identical in construction to CHAIR-S, and cover is the micro fraction of GT objects mentioned, identical in construction to CHAIR recall. Object factuality is $1 - \text{hal}$ and generation coverage is cover; the reported T4 values (Table 21) are read from a single frozen factuality/coverage pack.

V Mezzanine Measured Updates

This section collects the verification-frozen incremental values that were measured after the primary tables closed, each with its artifact provenance.

Statistical-power disclosure. The residual-risk denominator at the frozen operating point is small (8 hallucinated mentions of 751 candidate mentions across 500 evaluation images), so the bootstrap interval degenerates to the point estimate. We therefore report the pre-registered verdict as mechanistic (CI lower $> \delta$ holds degenerately) and complement it with the occurrence-level counts above, rather than interpreting the interval as informative. The guarantee fails on every audited occurrence (8/8); the absolute number of such cases is governed by the detector-side evidence coverage of the evaluation pool.

Fidelity note. The frozen headline revision-rate pair is stated in Sec. 7; this note discloses the implementation-level gap behind it. The official selection implementation is not public; under the frozen paper-specified quantities, our selection distribution is dominated by sample candidates chosen by sequence log-probability flips (478/500 sample, 19 beam, 3 greedy; 462/500 logp-flip-driven). We disclose this gap rather than tune toward the reference value, because the guarantee analysis (GT sensitivity via $\hat{R}$, τ, and the pre-registered break rule) is invariant to the absolute revision rate.

Table 20: AMBER generative (frozen assets, n=1004), LLaVA-1.5-7B, the pinned llava-hf revision. The headline CHAIR value is the caption-level CHAIR analog hal used throughout this paper; under the strict official object-level convention, the corresponding value for the base model is 0.0084. Metric definitions are collected in App. U.

model	CHAIR ↓	cover ↑	cog ↓	CHAIR 95% CI
LLaVA-1.5-7B	0.1016	0.3944	0.0275	[0.084, 0.122]

Table 21: T4: independent-generation factuality on the test1000 convention; the holdout view is Figure 17. These captions are scored as generations and never rebased into the same-caption edit metric. The factuality/coverage columns apply the AMBER hal/cover convention transferred to the COCO-GT setting; the transfer is provisional, pending upstream adjudication. Bold marks the best cell per column; all ties are bold. Registered as T4 in Table 17. Metric definitions are collected in App. U.

Generator	CHAIR-S ↓	CHAIR-I ↓	Object factuality ↑	Generation coverage ↑
Base VLM	0.5000	0.1470	0.5000	**0.7954**
REVERSE	0.4680	**0.1300**	0.5320	0.7735
VCD	0.5170	0.1663	0.4830	0.7822
OPERA	**0.2110**	0.1919	**0.7890**	0.5187
HALC	0.5030	0.1455	0.4970	0.7934

W Restoring C/U on Detector-Cache Columns

Why C/U need restoring here. Some rows of the T2 mechanism table dissect why detection works, and those rows need the generator's own confidence C and uncertainty U per claim. The reverse-uncertainty row reads U and the legacy three-term risk row reads both (Eq. (1)). LLaVA-1.5-7B has both signals measured from recorded token streams. The four detector-cache columns — cached detector outputs over generations whose token traces were never saved — have neither. Rather than dropping those columns' C/U-dependent rows or fabricating values, we fill them under three conventions, all reported in Table 19.

The three conventions. Case A, the subject of this section, restores C/U from the measured token stream of the same caption. Case B leaves the C/U-dependent rows n/a on detector-cache columns by design. Case C is the dev300 rebuild from the post-fix span-extraction pipeline.

How Case A is built. On a restored column, which carries no measured C/U by design, the C/U-dependent T2 rows are rebuilt from the token-level logprob and entropy exports of the same caption: span-level confidence is the geometric mean of the span's token probabilities (Eq. (1)) and span-level entropy is the mean token entropy. The restored confidence is used as C and the restored entropy as U. The restoration joins spans by image and text in first-in-first-out occurrence order, and the join coverage is 100% on every reported column.

Scope and honesty. Case-A restoration exists only for the Qwen3-VL 8B and 235B case-A blocks and the case-C dev300 row; gemma-4-31B-it exported no token stream and stays case B. The restored values keep the semantics of the measured token distribution, and differences versus the LLaVA column's directly measured C/U are disclosed rather than equated.

X Generation Factuality, Standard Benchmarks, and Cross-Model Replication

Tables 21, 22, and 23 collect the measured tracks held in this appendix: independent-generation factuality, standard hallucination benchmarks, and cross-model replication, each with its own caption.

Table 22: T5: standard hallucination benchmarks. Every row requires official frozen method revisions and a complete output set. The MME (Fu et al., 2023) column is omitted rather than marked n/a by registered decision: it was descoped from the frozen benchmark set and is carried as a declared omission for camera-ready follow-up; no values are fabricated. POPE cells list the random, popular, and adversarial settings in that order. Bold marks the best cell per column, comparing the POPE triplet by its mean across the three settings; all ties are bold. Registered as T5 in Table 17. Metric definitions are collected in App. U.

Method	CHAIR-S ↓	POPE F1 ↑	AMBER Hall. ↓	AMBER Cover. ↑
Base VLM	0.5000	0.886/0.857/0.803	0.1016	0.3944
REVERSE	0.4680	**0.903/0.887/0.841**	0.1066	**0.4281**
VCD	0.5170	0.870/0.838/0.794	0.1135	0.4009
OPERA	**0.2110**	0.885/0.856/0.803	0.1255	0.3676
HALC	0.5030	0.886/0.857/0.803	0.1056	0.3930
Woodpecker	0.5010	0.733/0.724/0.718	**0.0956**	0.3887
EVIC-VL	0.5040	0.887/0.875/0.849	0.1016	0.3944

Table 23: T6: cross-model replication over at least two frozen VLM revisions. The headline is the paired ΔAUROC contrast only; the human Edit F1 cells remain descoped for the cross-model track: T3 (Sec. 7) reports Edit F1 on the primary generator only, so no cross-model Edit F1 numbers are reported. The ΔAUROC cells are the dev300 pilot under column-native labels (row A: LLaVA−gemma; row B: its exact mirror); the intervals touch zero, so the difference is not resolvable at pilot n Per-column AUROC support is in App. O; the pilot contrast is the plug-in estimate, whereas App. U uses the replicate median, ± 0.067. CHAIR-S and POPE F1 columns are the frozen T5 generation-track values for context; POPE cells list the three settings in the App. U order. Registered as T6 in Table 17. Metric definitions are collected in App. U.

Frozen VLM revision	ΔAUROC [95% CI] ↑	CHAIR-S ↓	POPE F1 ↑
LLaVA-1.5-7B	-0.074 [-0.222, 0.000]	0.5000	0.886/0.857/0.803
gemma-4-31B-it	+0.074 [0.000, +0.222]	0.3130	0.888/0.870/0.848

Y LABEL AUDIT

Because COCO annotations can omit visible objects, the CVH label of Equation (2) can err in one direction. We therefore audit the labels directly with method-independent human raters.

Population and sampling. The audit population comprises the 539 CVH candidate claims flagged by the rule across the three splits — 53 on dev, 188 on test, and 298 on holdout — spanning 644 underlying occurrences, identified by the Stage-1 detector triage (App. H) that re-checks annotation absence with Grounding DINO caches. The audit packet samples these into 354 distinct (image, object) rows (spanning 437 of the 644 occurrences), drawn from the 18 most frequent canonical categories, with additional COCO-present controls drawn from the same images; the controls act as false-alarm probes and never enter the precision denominator. Detachment from the method is structural: the rater sees only the image and the object name, never the stratum, detector scores, or the model caption, and the display order is re-shuffled per rater, seeded by the rater identifier.

Protocol and reliability. The audit is two-rater: one of the authors rates the full frozen audit packet (354 candidate rows + 441 control probes), and a second blinded rater covers a 160-row stratified subset. Both raters use the same pre-specified judgment vocabulary (present / absent / unresolvable), with unresolvable items excluded from rate denominators and counted (7 for the author, 8 for the second rater). Reliability is disclosed through three measurable signals: the control false-alarm rate, the author's self-agreement on a 20% washout re-rated subset, and Cohen's κ between the two raters on the 160-row subset.CVH label precision is the share of audited CVH candidates confirmed truly absent; its interval follows Equation (19). Pooled over both audited strata, the rater confirmed 0.831 of rated candidate rows as truly absent (95% CI [0.788, 0.867]); the headline precision is reported

Table 24: T7: label-quality audit of the CVH candidate rule. The Wilson interval is over the correctness of candidates retained after detector triage. Human rows are computed from the author rater's judgments and the machine rows report the frozen rule, so the table reads as one evaluation chain. Registered as T7 in Table 17. Metric definitions are collected in App. U.

Quantity	Source	Value
Frozen candidate rule (machine)		
CVH candidate claims flagged	frozen rule, all splits dev 53 / test 188 / holdout 298	539
Blinded audit (human)		
Audited candidates	author rater	354
Control probes	author rater	441
CVH candidate precision	rater, on triage-retained candidates (n=196)	0.918
Wilson 95% CI	rater	[0.872, 0.949]
Demoted candidates confirmed present	rater, on triage-demoted candidates (n=153)	0.281 [0.216, 0.357]
Control false-alarm rate	control probes	0.182
Re-rated self-agreement	20% washout re-rating	0.925
Second-rater subset	second rater	160
Cohen's κ, author vs. second rater	subset	0.497

Table 25: Post-audit sensitivity of the T1 contrast, run as a sensitivity analysis rather than the primary endpoint. Labels are re-derived after the Stage-1 detector-triage audit; the annotation-defined labels remain the primary population (Sec. 3, audit independence). All three splits pass the registered gate under audited labels. The post-audit significance analysis recomputes the primary contrast on audited labels with the same image-cluster bootstrap (n=3000, seed 0) as Table 4. Metric definitions are collected in App. U.

Split	OWLv2	Ensemble	ΔAUROC [95% CI]	n obj / CVH
dev300	0.8720	0.9626	+0.0897 [+0.0515, +0.1345]	423 / 30
test1000	0.8951	0.9622	+0.0668 [+0.0475, +0.0879]	1446 / 90
holdout2100	0.9224	0.9792	+0.0565 [+0.0448, +0.0697]	2,868 / 160

on the triage-retained stratum by the frozen design. The machine rule and human audit rows are collected in Table 24.

Z Claim–Evidence Map

Table 26 lists, for every claim class in the paper, the minimum evidence and the prohibited shortcut.

Table 26: Minimum evidence required for each paper claim. A failed registered test changes the conclusion; it does not authorize a replacement endpoint.

Claim class	Required evidence	Prohibited shortcut
Ensemble improves detection	T1 effect, clustered CI, SESOI gate	validation/test reuse or row bootstrap
Fixed-claim uncertainty blind spot	T2 identical-claim scores	comparing independent captions
Editing effectiveness (T3)	executed blinded study (105 items, two raters, adjudicated), Edit F1–retention pair	claiming edit gains beyond the adjudicated study (e.g., superiority despite overlapping intervals)
Generation factuality	T4/T5 complete official outputs	rebasing generations into edits
Cross-model replication	T6 at least two frozen revisions	swapping only prompts or seeds
Benchmark label quality	T7 blinded audit + control probes	method detector as ground truth
Component contribution	T8 behaviorally distinct ablations	toggling unwired configuration flags