EViCA: Evidence-Guided Visual Claim Adjudication for Mitigating Hallucinations in Large Vision-Language Models

Anonymous authors
Paper under double-blind review

Abstract

Tool-assisted correction of large vision-language models requires evidence that is relevant to the target claim and sufficient to justify a change. A missing detection alone does not establish that an object is absent, and each revision must remain within the scope justified by its evidence. We introduce EViCA, a training-free framework for evidence-guided visual claim adjudication and selective correction. EViCA associates target claims with object-level observations and their provenance, then applies programmatic checks to assess claim-specific conditions for support and refutation. Unmet conditions guide additional observations within a fixed budget before final adjudication. Task-specific intervention rules govern answer corrections and constrain caption edits to eligible object mentions whose existence claims are refuted. For unresolved or conflicting claims, the controller retains the original answer or object mention. We evaluate EViCA on visual question answering and image captioning, examining correction quality, harmful modifications, and content retention.

1 Introduction

Large vision-language models (LVLMs) answer visual questions and describe images, but their responses can assert absent objects, miss present objects, or misstate their attributes (Li et al., 2023; Rohrbach et al., 2018; Fu et al., 2025). Tool-assisted correction inspects the image after generation to revise these errors without retraining the LVLM. Because interventions can also replace correct answers or remove true descriptions, correction must determine which claims warrant a change and how much of the response to edit.

Contrastive decoding strengthens image-conditioned predictions, and localized grounding helps resolve uncertain details during generation (Leng et al., 2024; Favero et al., 2024; Chen et al., 2024). After generation, Woodpecker organizes visual checks before revision, while Pelican decomposes claims, executes verification programs, and uses an LLM to assess the resulting evidence (Yin et al., 2024; Sahu et al., 2024). Both evidence acquisition and its interpretation therefore shape the correction of an existing response.

Our diagnostics expose three failures in this decision process (Figure 1). A skateboard detection can incorrectly reverse an answer about a baseball bat. An empty tool response can trigger a negative answer despite the target being present. A model-based rewrite can remove a true object mention while correcting a false one. The first two failures concern how observations establish a claim's status; the third concerns how that status is translated into an edit. Together, they motivate an explicit link between evidence, claim judgment, and permitted changes.

We introduce **EViCA**, a training-free framework for evidence-guided visual claim adjudication and selective correction. Its central design is to use claim-specific conditions to govern both evidence acquisition and intervention. Support requires observations that match the target and bind its predicates to compatible object instances. Refutation requires evidence against the claim together with checks that make this evidence applicable; for example, an absence decision must account for the coverage and completion of the search. Unmet conditions direct further observations within a fixed budget.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

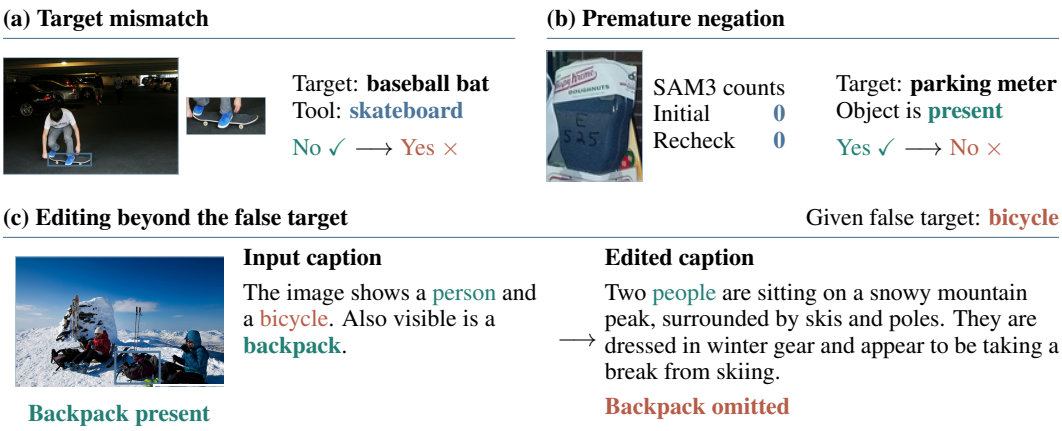

Figure 1: **Three failures in visual correction.** (a) Evidence concerns the wrong target. (b) Empty observations trigger a false negation. (c) Editing the supplied false target, bicycle, also removes the true backpack mention. Diagnostic protocols and full images appear in Appendix B.

EViCA makes the decision to intervene a programmatic step. It evaluates support and refutation separately, so a claim can remain unresolved or have conflicting evidence. Task-specific rules then translate the result into an answer correction or a local caption edit. Unresolved and conflicting claims retain their corresponding answer or object mention. Caption edits are restricted to eligible refuted mentions and checked against the remaining object inventory.

Our contributions are as follows:

- We identify target mismatch, premature negation, and excessive editing in tool-assisted visual correction. Controlled diagnostics show how each failure motivates a requirement on claim alignment, evidence sufficiency, or edit scope.
- We propose EViCA, which uses explicit support and refutation conditions to guide further inspection and selective correction. The framework links instance-associated visual evidence to answer changes and constrained caption edits without retraining the LVLM.
- Experiments on visual question answering and image captioning show improved factuality across two LVLMs. Controlled ablations isolate how claim adjudication reduces harmful changes, while acquisition and caption analyses reveal the trade-offs in observation requests and content retention.

2 Related Work

Interventions during generation. Generation-time methods change visual conditioning or token selection before a response is complete. VCD and M3ID contrast image-conditioned predictions with distorted-image or language-only references (Leng et al., 2024; Favero et al., 2024). SID, AvisC, ONLY, and DiVE construct decoding references through visual-token or internal model interventions (Huo et al., 2025; Woo et al., 2025; Wan et al., 2025; Li et al., 2026). OPERA and SGRS/LocoRE target unreliable attention or contextual dependencies during decoding (Huang et al., 2024; Zhang et al., 2026). Other approaches intervene on image regions or visual representations: HALC uses local focal contrast, AIR reinforces selected patches, and SHIELD addresses encoder bias and vulnerability (Chen et al., 2024; Zhu et al., 2026; Huang et al., 2026). EViCA operates on completed responses, where the decision concerns which existing claims to revise and which to preserve.

Claim verification and correction. Post-generation correction can learn a revision policy, as in LURE (Zhou et al., 2024), or use an explicit verification pipeline, as in Woodpecker (Yin et al., 2024). Visual programming composes perception modules into executable reasoning procedures (Gupta & Kembhavi, 2023; Surís et al., 2023); Pelican applies this idea to claim predicates and instance variables, then uses an LLM to assess the collected evidence (Sahu et al., 2024). Evidence

acquisition itself can also be selective: BCEA combines budgeted acquisition with conformal prediction, while Active-Look uses grounding-expert disagreement to guide contextual highlighting and zooming (Xu et al., 2026; Jiang et al., 2026). EViCA centers the decision procedure on executable support and refutation conditions. These conditions identify what further evidence is needed and govern whether the existing output can be changed, with separate controls for answer revision and caption editing.

Evaluating visual factuality. CHAIR measures hallucinated object mentions, while POPE and MME probe existence and other visual properties through questions (Rohrbach et al., 2018; Li et al., 2023; Fu et al., 2025). HallusionBench distinguishes language hallucination from visual illusion using controlled image–question pairs (Guan et al., 2024), and FaithScore evaluates descriptions through atomic visual facts (Jing et al., 2024). For correction, the transition from the original output to the revised one is also informative. We therefore pair benchmark scores with beneficial and harmful answer changes, and examine how caption editing affects the retention of true object content.

3 Preliminaries and Motivation

3.1 Problem Setup

Given an image I and a prompt q, a frozen LVLM produces an initial output y_0. We study two forms of correction: revising binary answers to visual questions and removing false object mentions from image captions. A target claim c is the proposition expressed by a question or the existence claim associated with a caption mention. A visual tool returns observations $\mathcal{E}$, including candidate regions, masks, scores, and the queries that produced them. Within an observation budget B, the system uses this evidence to produce a corrected output y. Its objective is to resolve erroneous claims while preserving correct answers and useful caption content.

3.2 Claim–Evidence Alignment

Figure 1(a) shows a target mismatch: a skateboard detection reverses a correct No to a question about a baseball bat. We study this failure in a Qwen2.5-VL-7B and SAM 3 pipeline on 120 COCO images, using COCO positive annotations and LVIS explicit negatives to construct 480 balanced existence questions (Lin et al., 2014; Gupta et al., 2019). A label-independent rule selects one recorded decision per question. Alternative rules receive the same inputs and observations; Appendix B gives the protocol and confidence intervals.

Let G and H denote wrong-to-correct and correct-to-wrong changes relative to the answer immediately before the selected decision. Checking target correspondence reduces H from 20 to 4 while retaining 31 of the 34 beneficial changes (Table 1). Thus, even when a tool returns a valid object detection, its relevance to the claim is a consequential part of correction.

Table 1: **Local decision diagnostics.** The 480 inputs are fixed; net correction is $G - H$.

Decision rule	$G \uparrow$	$H \downarrow$	$G - H \uparrow$
Original rule	34	20	14
Target correspondence	31	4	27
Preserve on empty evidence	34	15	19
Both checks	31	0	31

For compositional claims, alignment also requires predicates about one entity to refer to a compatible object instance (Ma et al., 2023). Separate observations for predicates p_1 and p_2 do not establish their conjunction on the same object:

$$(\exists x_1\, p_1(x_1)) \wedge (\exists x_2\, p_2(x_2)) \;\not\Rightarrow\; \exists x\, [p_1(x) \wedge p_2(x)]. \tag{1}$$

This stronger requirement motivates retaining instance associations alongside the target category. Section 4 formalizes it through compatible assignments of claim entities to observed objects.

3.3 Evidence Sufficiency

An empty result describes what a tool found under one query and view. For an existence claim c, let $\mathcal{O}(c, I, a)$ be the target-matched candidates returned under acquisition action a. Then

$$\mathcal{O}(c, I, a) = \varnothing \;\not\Rightarrow\; I \models \neg c. \tag{2}$$

Object scale, occlusion, or limited search coverage can produce an empty result even when c is true. A matching witness can establish existence; an absence decision also depends on whether the observation procedure covered the relevant regions and completed the required checks.

The diagnostic data contain 261 initial observations with no retained candidates, of which 21 concern objects annotated as present. Preserving an affirmative answer when the recorded observations provide no positive witness prevents five harmful local changes. Combined with target correspondence, this rule yields 17 additional correct decisions over the original rule (Table 1). These fixed-observation comparisons show why a correction policy needs a criterion for refutation beyond the failure to find support.

We next test whether further inspection can resolve missing evidence. Branches starting from the same initial observation use an unchanged retry, input resampling, local tiles, or candidate crops. Across branches, 18 image–target pairs gain a newly matched instance; five gain their first annotation-matched witness. The unchanged retry adds none. A local tile also returns a candidate for one explicitly absent target. Additional observations can therefore provide both new witnesses and false candidates. EViCA directs further inspection toward unmet claim conditions and checks the accumulated evidence before intervention.

3.4 Intervention Scope

Figure 1(c) shows why edit scope matters after adjudication. Given bicycle as the false target, the editor removes the bicycle mention but also omits the true backpack mention. Successful correction of the designated target can therefore coincide with loss of other correct content.

We isolate this effect on 48 images using two caption templates, each containing two annotated true objects and one explicitly absent object. Two paired repeats yield 192 outputs per editor. With the false target supplied, the model-based editor resolves it in all 192 outputs. A contextual review identifies loss of another true fact in 61 outputs (31.77%); 37 outputs remain ambiguous. A span-restricted editing probe preserves both specified true facts in all 192 outputs of these templates. This comparison motivates a separate constraint on intervention: retain the text span associated with each claim, edit only eligible refuted mentions, and check the complete object inventory after the edits are combined.

4 Proposed Method

4.1 Overview

Figure 2 summarizes EViCA. Starting from y_0 (R_0 in the figure), we derive question or caption claims and acquire SAM 3 observations. We test joint support through instance-consistent witnesses and check refutation through conditions specific to each type of contrary evidence. Unmet conditions guide further observations within a budget. Final adjudication then supplies the evidence state to the QA or caption controller, which determines the allowed correction. Model weights remain unchanged. Algorithm 1 gives the complete procedure.

4.2 Instance-Consistent Witnesses

Combining observations requires deciding which ones describe the same object. We map SAM 3 (Carion et al., 2025) boxes and masks to the original image, retaining each observation's query, score, view, action, and round. The existence module links observations through Same, Different, or Ambiguous relations. Candidates connected by Same relations form an entity group. Each group contributes its maximum valid support score, so another view can strengthen support without being counted as another object.

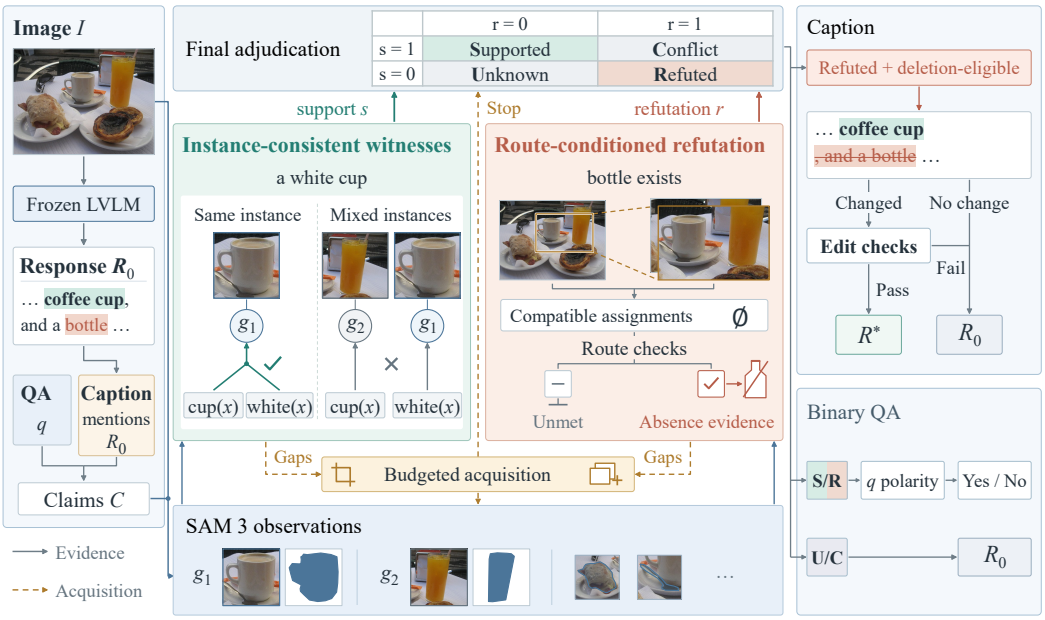

Figure 2: EViCA framework. Claim-specific support and refutation checks guide acquisition and selective correction. The photo and caption edit use COCO 135029. Central examples are schematic, and COCO annotation masks illustrate the observation representation.

Let $\mathcal{L}_c$ be the claim's required conditions, or *literals*, and $\mathcal{A}_c(\mathcal{E})$ the compatible entity assignments examined by the judge. Literals sharing a variable bind it to the same entity. Given valid literal support $u(\ell, a; \mathcal{E})$ under assignment a and threshold τ_s, we compute

$$\gamma(c; \mathcal{E}) = \max_{a \in \mathcal{A}_c(\mathcal{E})} \min_{\ell \in \mathcal{L}_c} u(\ell, a; \mathcal{E}), \qquad s = \mathbf{1}[\gamma(c; \mathcal{E}) \geq \tau_s], \tag{3}$$

The minimum measures the weakest required literal within one assignment; the maximum selects the strongest compatible witness. An empty assignment set has zero support.

Strong individual cues can still fail to form a joint witness. To identify this case, let $\mathcal{E}_\ell^+$ contain valid positive evidence for literal ℓ, with score $\sigma(e)$. When every such set is nonempty, define

$$\beta(c; \mathcal{E}) = \min_{\ell \in \mathcal{L}_c} \max_{e \in \mathcal{E}_\ell^+} \sigma(e), \qquad \Delta_{\text{bind}} = \max\{0, \beta(c; \mathcal{E}) - \gamma(c; \mathcal{E})\}. \tag{4}$$

The marginal score β selects each literal's evidence independently. The binding gap measures the loss from requiring a compatible assignment. For an Unknown claim, $\Delta_{\text{bind}} \geq 0.15$ marks a need for instance disambiguation during acquisition. If any $\mathcal{E}_\ell^+$ is empty, the gap is zero. For POPE's single-category claims, support is the score of the strongest valid matching candidate. Count, position, and color use their own typed operators.

4.3 Route-Conditioned Refutation

Low support alone cannot distinguish a false claim from missing evidence. A *refutation route* ρ therefore combines an evidence condition D_ρ directed against the claim with validity conditions $\mathcal{K}_\rho$ that determine whether this evidence covers the conclusion's scope:

$$r = \mathbf{1}\left[\exists \rho : \; D_\rho(c, \mathcal{E}) \; \wedge \bigwedge_{k \in \mathcal{K}_\rho} v_{\rho,k}(c, \mathcal{E})\right]. \tag{5}$$

Here $v_{\rho,k}$ indicates whether condition k passes. Refutation requires at least one route whose evidence condition and all associated checks pass.

For object absence, D_ρ requires an empty compatible target-assignment set. The accompanying checks cover the whole image, required scales, concept queries, image quality, completion of acquisition requirements, and tool health. Truncation or unfinished requirements leave this route incomplete. A nonempty set of candidates below τ_s establishes neither support nor absence: it supplies no strong witness but blocks the empty-set condition.

These requirements depend on the claim. Four distinct dogs refute an exact count of three, whereas establishing an exact count requires a completed candidate domain. Position compares ranges of possible center coordinates for bound objects, and color measures their mask pixels using HSV and CSS-Lab. These typed operators check evidence completeness and conclusion stability across ordered score thresholds or mask-erosion levels (Appendix A.2).

4.4 Budgeted Evidence Acquisition

Unmet conditions indicate what another observation should resolve. The planner requests resized views or overlapping tiles for missing scale coverage, and targeted crops or renewed observations for ambiguous instances. We remove duplicate requests and select actions by fixed priority, remaining budget, and action caps. The shared existence pipeline allows an initial observation and at most two refinement rounds, with separate action caps.

Acquisition criteria guide further observations; final adjudication criteria determine whether the accumulated evidence permits a correction. When no requested action fits the remaining budget, acquisition stops and the final judge evaluates the accumulated record once. It uses recorded completion, tool failures, and truncation to identify unmet requirements without further tool calls. Unfinished refutation conditions remain unresolved. Appendix A.1 details this execution order.

4.5 Selective Correction

Separate support and refutation checks distinguish missing evidence from contradictory evidence:

$$\mathrm{state}(c;\mathcal{E}) = \begin{cases} \text{Supported} & (s,r) = (1,0), \\ \text{Refuted} & (s,r) = (0,1), \\ \text{Unknown} & (s,r) = (0,0), \\ \text{Conflict} & (s,r) = (1,1). \end{cases} \tag{6}$$

For binary QA, the controller preserves y_0 when the state is Unknown or Conflict, or when the decision fails the task's output checks. A permitted Supported or Refuted state maps to Yes or No through question polarity. MME count and position require complete, stable decisions, with an additional count criterion described in Appendix A.2. LLaVA's color policy retains the original answer.

For captions, each deletion is tied to the source span of a Refuted existence claim. The span must be eligible for removal and accompanied by a refutation record bound to that claim. Other evidence states preserve the corresponding mention while eligible Refuted mentions remain editable. We combine accepted deletions with local grammatical repair, then extract the object inventory again. The combined edit must satisfy an inventory constraint derived from retained mentions, remain nonempty, and pass local grammatical checks. A failed check restores y_0. Appendix A.3 gives the deletion criteria.

5 Experiments

5.1 Experimental Setup

We evaluate EViCA with Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and LLaVA-v1.5-7B (Liu et al., 2024) under standard benchmark protocols. Both backbones remain frozen, with SAM 3 supplying visual evidence. The final existence-support thresholds are 0.50 for Qwen and 0.70 for LLaVA. Captioning uses the same thresholds without CHAIR-specific calibration.

Benchmarks and selection. POPE (Li et al., 2023) evaluates object existence through 27,000 questions across MS-COCO, A-OKVQA, and GQA, using random, popular, and adversarial neg-

ative sampling. Main comparisons cover the complete benchmark. Qwen ablations exclude its 148 calibration images (2,682 questions), leaving 24,318 questions from 1,341 images. MME-Hallucination (Fu et al., 2025) extends the evaluation to existence, count, position, and color. Its 240 questions serve for both count-threshold selection and evaluation. For captioning, we sample 500 MS-COCO validation images (Lin et al., 2014) and repeat caption generation and correction three times, pairing each initial caption with its edited output and reporting mean CHAIR_S, CHAIR_I (Rohrbach et al., 2018), and object recall.

Comparison methods and metrics. We compare with VCD (Leng et al., 2024), M3ID (Favero et al., 2024), SID (Huo et al., 2025), AvisC (Woo et al., 2025), ONLY (Wan et al., 2025), and DiVE (Li et al., 2026). POPE Average takes the mean across sampling strategies within each source and then across sources. For EViCA's interventions, G counts incorrect initial answers corrected, H counts correct initial answers made incorrect, and $G - H$ measures net correction. Appendix C defines the metrics and calibration procedure.

5.2 Comparison with Existing Methods

Object existence. EViCA achieves the highest average accuracy and F1 on both backbones, improving accuracy by 6.30 and 10.18 percentage points over Base (Tables 2 and 3). Gains cover all three image sources and sampling strategies. On the held-out Qwen set, EViCA corrects false-negative answers about present objects (1,342 net corrections) and false-positive answers about absent objects (160 net corrections) in paired comparisons. Both directions contribute, with false negatives accounting for most of the net improvement in this subset.

Table 2: **Performance with Qwen2.5-VL-7B.** POPE/CHAIR are in percent; best results are bolded.

Method	POPE (Average)		CHAIR			MME
	Acc.↑	F1↑	$C_S \downarrow$	$C_I \downarrow$	Recall↑	Total↑
Base (Bai et al., 2025)	83.93	82.54	37.60	9.30	71.70	690.00
VCD (Leng et al., 2024)	84.86	83.26	37.80	9.40	70.80	695.00
M3ID (Favero et al., 2024)	84.51	83.15	33.80	9.20	69.30	700.00
SID (Huo et al., 2025)	84.99	83.67	36.20	9.10	72.50	703.30
AvisC (Woo et al., 2025)	85.70	84.95	35.40	8.90	70.70	708.30
DiVE (Li et al., 2026)	86.70	86.38	31.40	7.30	**73.00**	720.00
EViCA	**90.23**	**89.88**	**25.27**	**6.74**	65.87	**740.00**

Table 3: **Performance with LLaVA-v1.5-7B.** POPE/CHAIR are in percent; best results are bolded.

Method	POPE (Average)		CHAIR			MME
	Acc.↑	F1↑	$C_S \downarrow$	$C_I \downarrow$	Recall↑	Total↑
Base (Liu et al., 2024)	80.01	82.31	53.20	14.70	80.50	643.30
VCD (Leng et al., 2024)	78.72	81.39	51.20	14.10	79.70	648.30
M3ID (Favero et al., 2024)	78.84	81.88	54.60	14.80	**82.00**	653.30
SID (Huo et al., 2025)	80.56	83.19	50.20	13.30	78.90	650.00
ONLY (Wan et al., 2025)	79.89	82.63	50.80	14.50	80.20	653.30
AvisC (Woo et al., 2025)	81.93	82.98	51.60	13.80	79.30	665.00
DiVE (Li et al., 2026)	85.49	85.04	47.40	13.00	81.50	671.70
EViCA	**90.19**	**89.89**	**35.87**	**8.89**	75.82	**695.00**

Typed visual questions. EViCA reaches 740.00 on Qwen and 695.00 on LLaVA, exceeding Base by 50.00 and 51.70 points. Count and position contribute the largest gains (Table 12), consistent with the use of instance and spatial evidence for these claims. LLaVA's color score is 15.00 points below Base, while its higher scores on existence, count, and position yield a higher total score.

Caption correction and retention. Captioning tests the balance between suppressing false object mentions and retaining useful content. EViCA obtains the lowest CHAIR_S and CHAIR_I for both backbones. Relative to Base, CHAIR_S falls by 12.33 and 17.33 percentage points for Qwen and LLaVA, and CHAIR_I by 2.56 and 5.81 points. Object recall also falls, to 65.87% and 75.82%. The mean ratios of edited to original token counts are 91.71% and 87.90%, respectively. Edited captions average 144.50 and 79.45 words, with no empty captions.

5.3 Ablation Studies

We examine additional observations, adjudication on identical evidence, and evidence-guided acquisition. All variants share initial Qwen responses. Table 4 reports results on the 24,318 held-out questions. Requests count logical observations, including reuse.

Table 4: **Component ablations on Qwen POPE.** Best accuracy and F1 are bolded.

Variant	Acc.	F1	G	H	Requests
Initial only	89.58	89.58	1,651	317	24,318
Score only	89.77	89.28	2,276	895	51,258
Fixed schedule	**90.30**	**89.99**	2,235	726	74,598
EViCA (Full)	90.27	89.93	2,193	691	51,258

Additional observations. Initial only isolates the contribution of additional observations by retaining Full's adjudicator while removing observations after the initial whole-image query. Full improves accuracy by 0.69 percentage points (95% paired image-bootstrap interval $[0.34, 1.03]$). Its beneficial and harmful change counts increase by 542 and 374, respectively, yielding 168 additional net corrections. Further evidence improves the aggregate result while introducing new errors.

Adjudication on identical evidence. Score only tests the decision rule on Full's acquired observations. It answers yes when maximum valid support reaches 0.50 and no otherwise, retaining Full's output in 60 cases flagged for image quality or search truncation. Full avoids 204 harmful changes while forgoing 83 beneficial ones, yielding 121 net corrections and a 0.50-point accuracy gain (95% interval $[0.29, 0.71]$). The two policies differ on 287 questions, all with nonempty evidence below the support threshold: Full preserves the original yes answer, whereas Score only changes it to no. The advantage comes from preserving the original answer when insufficient support does not establish absence.

Performance and observation requests. Fixed schedule uses the same action types, budgets, and adjudicator as Full, but draws requests from a fixed queue determined by the initial view and candidates. It obtains slightly higher accuracy and F1 ($90.30/89.99$) than Full ($90.27/89.93$), using 74,598 rather than 51,258 requests (Figure 3(a)). Full reduces the request count by 31.29%, with small observed differences in accuracy and F1. The paired intervals for both metric differences include zero (Table 14).

Evidence resolution. Adjudication states reveal how acquisition resolves claims. Initial only leaves 13,372 claims Unknown, whereas Full leaves 621 and yields 11,034 Supported and 12,663 Refuted claims (Table 16). Additional observations supply positive witnesses and satisfy refutation requirements. Fixed schedule resolves more claims, leaving 432 Unknown, consistent with its higher request count. None of these variants produces Conflict on this set. These counts measure evidence resolution: an Unknown claim can already have a correct initial answer that the controller preserves.

Threshold stability. The threshold sweeps test sensitivity around the selected cutoffs (Figure 3(b,c)). On the 2,682-question calibration set, Qwen accuracy changes by 0.04 percentage points between support thresholds 0.50 and 0.55, and LLaVA varies by 0.11 points over 0.65–0.75. Scores are locally stable, but higher thresholds reduce Qwen's net corrections as fewer positive answers pass the support gate (Table 8). Count qualification has a wider plateau: both backbones obtain a

count score of 180 at all four tested values from 0.65 to 0.80. The selection rule in Appendix C.2 chooses 0.80 within this plateau; raising it further reduces the score.

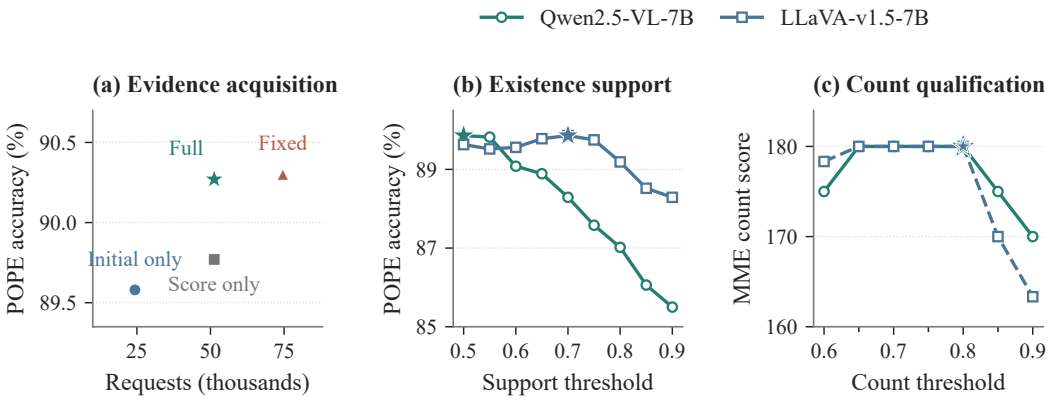

Figure 3: **Component and parameter analysis.** (a) Accuracy versus observation requests on Qwen POPE. (b,c) POPE and MME threshold sensitivity; stars mark the selected thresholds.

5.4 Computational Cost

Figure 4 compares GPU memory and average latency on Qwen. EViCA uses more memory and has higher latency than Base and the compared methods, alongside higher POPE accuracy and lower CHAIR scores. Evidence acquisition and adjudication dominate latency; claim processing and correction contribute little (Table 20). Table 21 reports the absolute comparison values.

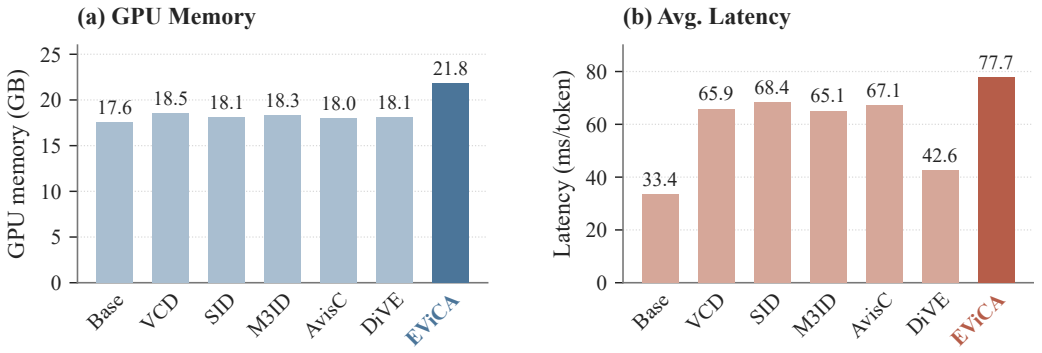

Figure 4: **Computational cost with Qwen2.5-VL-7B.** GPU memory and average latency. Memory is shown in GB.

6 Conclusion

EViCA uses explicit support and refutation conditions to guide evidence acquisition and selective correction without retraining the LVLM. Across two backbones, EViCA improves visual question answering and reduces object hallucination in captions. The matched-observation ablation shows that adjudication reduces harmful changes while retaining most beneficial corrections. Acquisition and caption analyses characterize the trade-offs between accuracy and observation requests, and between hallucination reduction and object recall.

AI USE STATEMENT

Generative AI assistants were used for literature discovery and analysis, research ideation, feedback on experimental design, implementation and analysis support, qualitative inspection of controlled editing outputs, figure preparation, and drafting and revising manuscript text. The authors are responsible for the scientific claims, experimental interpretation, and final submitted text.

ETHICS STATEMENT

The experiments use existing visual benchmarks and their annotations. The controlled caption study constructs object-existence statements from public dataset labels and evaluates model-generated revisions. It involves no recruited human participants. The study examines both correction benefits and the loss of previously stated visual content, which is relevant to using generated descriptions as faithful records of images.

REPRODUCIBILITY STATEMENT

We plan to open-source the Base reproduction code, benchmark configurations, and evaluation scripts for Qwen2.5-VL-7B-Instruct on POPE. EViCA will be provided through a class-based interface, with its core implementation remaining closed source. Section 4 and Appendix A describe the method and its pseudocode, while Appendix C specifies the evaluation protocols, thresholds, and metrics.

REFERENCES

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

Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python*. O'Reilly Media, 2009. URL https://www.nltk.org/book_1ed/.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, Jie Lei, Tengyu Ma, Baishan Guo, Arpit Kalla, Markus Marks, Joseph Greer, Meng Wang, Peize Sun, Roman Rädle, Triantafyllos Afouras, Effrosyni Mavroudi, Katherine Xu, Tsung-Han Wu, Yu Zhou, Liliane Momeni, Rishi Hazra, Shuangrui Ding, Sagar Vaze, Francois Porcher, Feng Li, Siyuan Li, Aishwarya Kamath, Ho Kei Cheng, Piotr Dollár, Nikhila Ravi, Kate Saenko, Pengchuan Zhang, and Christoph Feichtenhofer. SAM 3: Segment anything with concepts. *arXiv preprint arXiv:2511.16719*, 2025. URL https://arxiv.org/abs/2511.16719.

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, 2024. URL https://proceedings.mlr.press/v235/chen24bi.html.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multimodal hallucination control by visual information grounding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14303–14312, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Favero_Multi-Modal_Hallucination_Control_by_Visual_Information_Grounding_CVPR_2024_paper.html.

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. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4899. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/d79a27cf2772fe00be7f341efc0eb517-Abstract-Datasets_and_Benchmarks_Track.html.

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. URL https://openaccess.thecvf.com/content/CVPR2024/html/Guan_HallusionBench_An_Advanced_Diagnostic_Suite_for_Entangled_Language_Hallucination_and_CVPR_2024_paper.html.

Agrim Gupta, Piotr Dollár, and Ross Girshick. LVIS: A dataset for large vocabulary instance segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 5356–5364, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Gupta_LVIS_A_Dataset_for_Large_Vocabulary_Instance_Segmentation_CVPR_2019_paper.html.

Tanmay Gupta and Aniruddha Kembhavi. Visual programming: Compositional visual reasoning without training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14953–14962, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Gupta_Visual_Programming_Compositional_Visual_Reasoning_Without_Training_CVPR_2023_paper.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.

Yiyang Huang, Liang Shi, Yitian Zhang, Yi Xu, and Yun Fu. SHIELD: Suppressing hallucinations in LVLM encoders via bias and vulnerability defense. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/cae31fa913f140b3126036cc474fd461-Abstract-Conference.html.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 6700–6709, 2019. URL https://openaccess.thecvf.com/content_CVPR_2019/html/Hudson_GQA_A_New_Dataset_for_Real-World_Visual_Reasoning_and_Compositional_CVPR_2019_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 *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/3cc87f2bd3e3b4df8f9217326761c322-Abstract-Conference.html.

Yubo Jiang, Xin Yang, Abudukelimu Wuerkaixi, Zheming Yuan, Xuxin Cheng, Cao Liu, Ke Zeng, Fengying Xie, Zhiguo Jiang, and Haopeng Zhang. Global context or local detail? adaptive visual grounding for hallucination mitigation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 15133–15152, 2026. doi: 10.18653/v1/2026.findings-acl.745. URL https://aclanthology.org/2026.findings-acl.745/.

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, 2024. doi: 10.18653/v1/2024.findings-emnlp.290. URL https://aclanthology.org/2024.findings-emnlp.290/.

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.

Xinwei Li, Li Lin, Hui Jiao, Li Yao, Tien-Tsin Wong, and Hanqian Wu. DiVE: Decoupling intra-layer visual evidence for mitigating hallucinations in large vision-language models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 37552–37568, 2026. doi: 10.18653/v1/2026.acl-long.1742. URL https://aclanthology.org/2026.acl-long.1742/.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, 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, 2023. doi: 10.18653/v1/2023.emnlp-main.20. URL https://aclanthology.org/2023.emnlp-main.20/.

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 *Computer Vision – ECCV 2014*, pp. 740–755, 2014. doi: 10.1007/978-3-319-10602-1_48. URL https://link.springer.com/chapter/10.1007/978-3-319-10602-1_48.

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, 2024. URL https://openaccess.thecvf.com/content/CVPR2024/html/Liu_Improved_Baselines_with_Visual_Instruction_Tuning_CVPR_2024_paper.html.

Zixian Ma, Jerry Hong, Mustafa Omer Gul, Mona Gandhi, Irena Gao, and Ranjay Krishna. CREPE: Can vision-language foundation models reason compositionally? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10910–10921, 2023. URL https://openaccess.thecvf.com/content/CVPR2023/html/Ma_CREPE_Can_Vision-Language_Foundation_Models_Reason_Compositionally_CVPR_2023_paper.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, 2018. doi: 10.18653/v1/D18-1437. URL https://aclanthology.org/D18-1437/.

Pritish Sahu, Karan Sikka, and Ajay Divakaran. Pelican: Correcting hallucination in vision-LLMs via claim decomposition and program of thought verification. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 8228–8248, 2024. doi: 10.18653/v1/2024.emnlp-main.470. URL https://aclanthology.org/2024.emnlp-main.470/.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision – ECCV 2022*, pp. 146–162, 2022. doi: 10.1007/978-3-031-20074-8_9. URL https://link.springer.com/chapter/10.1007/978-3-031-20074-8_9.

Dídac Surís, Sachit Menon, and Carl Vondrick. ViperGPT: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 11888–11898, 2023. URL https://openaccess.thecvf.com/content/ICCV2023/html/Suris_ViperGPT_Visual_Inference_via_Python_Execution_for_Reasoning_ICCV_2023_paper.html.

Zifu Wan, Ce Zhang, Silong Yong, Martin Q. Ma, Simon Stepputtis, Louis-Philippe Morency, Deva Ramanan, Katia Sycara, and Yaqi Xie. ONLY: One-layer intervention sufficiently mitigates hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3225–3234, 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Wan_ONLY_One-Layer_Intervention_Sufficiently_Mitigates_Hallucinations_in_Large_Vision-Language_Models_ICCV_2025_paper.html.

Sangmin Woo, Donguk Kim, Jaehyuk Jang, Yubin Choi, and Changick Kim. Don't miss the forest for the trees: Attentional vision calibration for large vision language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 1927–1951, 2025. doi: 10.18653/v1/2025.findings-acl.99. URL https://aclanthology.org/2025.findings-acl.99/.

Jian Xu, Yanning Wu, Delu Zeng, John Paisley, and Qibin Zhao. Look again before you abstain: Budgeted conformal evidence acquisition for reliable vision-language model. *arXiv preprint arXiv:2606.16667*, 2026. URL https://arxiv.org/abs/2606.16667.

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://link.springer.com/article/10.1007/s11432-024-4251-x.

Xiaofeng Zhang, Yuanchao Zhu, Chaochen Gu, Xiaosong Yuan, Qiyan Zhao, Jiawei Cao, Barrett Tang, Sinan Fan, Yaomin Shen, Chen Shen, and Hao Tang. Hallucination begins where saliency drops. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/196c8da9209b1977408d8771c4e7ee56-Abstract-Conference.html.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/fc625e831361cfcc82cb74224fdc66cb-Abstract-Conference.html.

Xingyu Zhu, Kesen Zhao, Liang Yi, Shuo Wang, Zhicai Wang, Beier Zhu, Hanwang Zhang, and Xiangnan He. Look carefully: Adaptive visual reinforcements in multimodal large language models for hallucination mitigation. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/83749a89d03df30c66af88020e4c7892-Abstract-Conference.html.

A Algorithmic Description

A.1 End-to-End Procedure

Algorithm 1 follows claims from observations to output decisions. For captions, the object-mention inventory constructed with NLTK (Bird et al., 2009) preserves exact text spans and deletion eligibility. The record $\mathcal{T}$ tracks completed actions, tool outcomes, and truncation. The set $\mathcal{H}$ denotes unmet evidence conditions.

Algorithm 1 EViCA: evidence-guided visual claim adjudication and selective correction

Require: Image I, task input q, task type t, frozen route policies and budget B
Ensure: Corrected output y and claim certificates $\{z_c\}$

$y_0 \leftarrow \text{LVLM}(I, q)$
$\mathcal{C} \leftarrow \textsc{CompileQA}(q)$ or $\textsc{CompileMentions}(y_0)$, according to t
for each claim $c \in \mathcal{C}$ **do**
  if c is an existence claim **then**
    $(\mathcal{E}, \mathcal{T}) \leftarrow \textsc{InitialSAM3}(I, c)$
    for each refinement round permitted by B **do**
      $G \leftarrow \textsc{AssociateInstances}(\mathcal{E})$
      $(s, r, \mathcal{H}) \leftarrow \textsc{CollectionJudge}(c, G, \mathcal{E}, \mathcal{T})$
      $\mathcal{A} \leftarrow \textsc{RegisteredActions}(\mathcal{H}, c, G)$
      $\mathcal{A} \leftarrow \textsc{DeduplicateAndSelect}(\mathcal{A}, \mathcal{T}, B)$
      if $\mathcal{A} = \varnothing$ **then break**
      $(\mathcal{E}, \mathcal{T}) \leftarrow \textsc{AcquireAndAccumulate}(I, c, \mathcal{A}, \mathcal{E}, \mathcal{T})$
    end for
    $\mathcal{R} \leftarrow \textsc{TerminalRequest}(c, \mathcal{E}, \mathcal{T})$
    $z_c \leftarrow \textsc{FinalJudgeOnce}(c, \mathcal{R})$
  else
    $D_c \leftarrow \textsc{TypedSAM3Domains}(I, c)$
    $z_c \leftarrow \textsc{TypedAdjudicate}(c, D_c)$
  end if
end for
$y \leftarrow y_0$
if t is QA **then**
  $y \leftarrow \textsc{QAControl}(y_0, \mathcal{C}, \{z_c\})$
else
  $\mathcal{S} \leftarrow \textsc{QualifyingRefutedSpans}(y_0, \mathcal{C}, \{z_c\})$
  $y' \leftarrow \textsc{DeleteAndRepair}(y_0, \mathcal{S})$
  $y \leftarrow y'$ if $\textsc{AggregateChecks}(y_0, y', \mathcal{S})$ pass; otherwise retain y_0
end if
return $y, \{z_c : c \in \mathcal{C}\}$

The existence branch relates observations to objects before evaluating support, refutation, and missing conditions. The planner regenerates requests from the current unmet conditions at each round, prioritizing closer views for ambiguous instances or broader coverage for absence. Requests are merged by action, query, target, and view, and previously completed requests are removed. Fixed priorities and action caps allow at most two refinement rounds after the initial observation.

At the end, candidates are filtered for final adjudication. Acquisition completion, tool health, and truncation determine which requirements remain unresolved. Finishing the allotted rounds does not itself complete an absence route. The judge evaluates the record once, without further observations, and returns a certificate z_c containing the claim's state and associated evidence.

A.2 Typed Adjudication

For MME, supplemental SAM 3 observations are associated into category-specific instance tracks, each representing an object across observations. Candidate domains contain the potentially relevant instances at each score threshold. The ordered thresholds form a score path. Count and position require complete evidence and validity checks for the selected conclusion, agreement at the last two thresholds, and no opposing determinate conclusion elsewhere on the path.

Distinct instances provide a count lower bound, which refutes an exact count when it exceeds that count. Supporting an exact count, or refuting it because too few instances are found, requires a closed candidate domain, meaning the counting route's required observations and validity checks are complete. After domain construction, the minimum score among observations retained at the candidate threshold determines whether a count conclusion qualifies to change the output. An established lower-bound refutation is exempt from this criterion.

Position compares possible center-coordinate intervals and uses the bindings required by containment and other structural relations. Front/behind remains Unknown. Color compares HSV and CSS-Lab measurements of the bound object's mask over ordered score thresholds and mask-erosion levels, separating complete certification from partial evidence. Section 5 and Appendix C.2 give the thresholds and task policies.

A.3 Output Control

For QA, the controller first checks whether a decision meets the task's conditions for changing the answer. These include the count and position requirements above and the LLaVA color rule, which retains the original answer. A permitted Supported or Refuted decision is mapped to Yes or No through question polarity. Unknown, Conflict, and decisions that fail these checks preserve y_0.

For captions, removal requires an eligible object span, a Refuted certificate with refutation score at least 0.80, completed evidence requirements, and a route or direct evidence record supporting refutation. For the SAM 3 existence route, completion of the absence conditions assigns a refutation score of 1. The certificate must identify the same claim and calibration policy, and the claim's object category must match the mention. Other evidence states preserve the corresponding mention, independently of deletions elsewhere. Accepted removals are combined through deterministic span deletion and local grammatical repair. When needed, an existence template lists eligible retained object mentions. The controller re-extracts object mentions and restricts their categories to those of retained mentions or their query terms, while checking nonempty output and grammatical conditions. If the combined edit fails a check, it restores y_0.

B Diagnostic Studies

B.1 Diagnostic Protocol

The diagnostic study in Section 3 uses a Qwen2.5-VL-7B-Instruct and SAM 3 correction pipeline, with the decision rules fixed before evaluation. We select 120 COCO val2014 images with LVIS metadata and screen them against recorded earlier evaluations using exact and perceptual near-duplicate checks. Figure 5 shows the complete cases summarized in Figure 1.

After category mapping and conflict removal, each image contributes two COCO-positive and two LVIS-explicit-negative existence questions, giving 240 positives and 240 negatives. Negative labels come exclusively from LVIS explicit negatives. Annotations are used for diagnostic evaluation and withheld from inference. This 480-question diagnostic set is separate from official POPE.

The pipeline allows at most three correction rounds. From its 1,093 recorded decisions, we select one per question using a fixed, label-independent rule: the first decision where the SAM signal and the answer after correction disagree, or the first decision if none disagree. Each alternative decision rule receives the same incoming answer and observations, and scoring stops at its immediate output. This paired design isolates the effect of the local rule from changes in later evidence acquisition.

Local G and H count incorrect-to-correct and correct-to-incorrect changes from the incoming answer A_t to the outgoing answer A'_t (Table 1). End-to-end counts compare the initial LVLM answer with the final run answer.

The 95% intervals use 5,000 image-cluster bootstrap draws, resampling all questions from each image together.

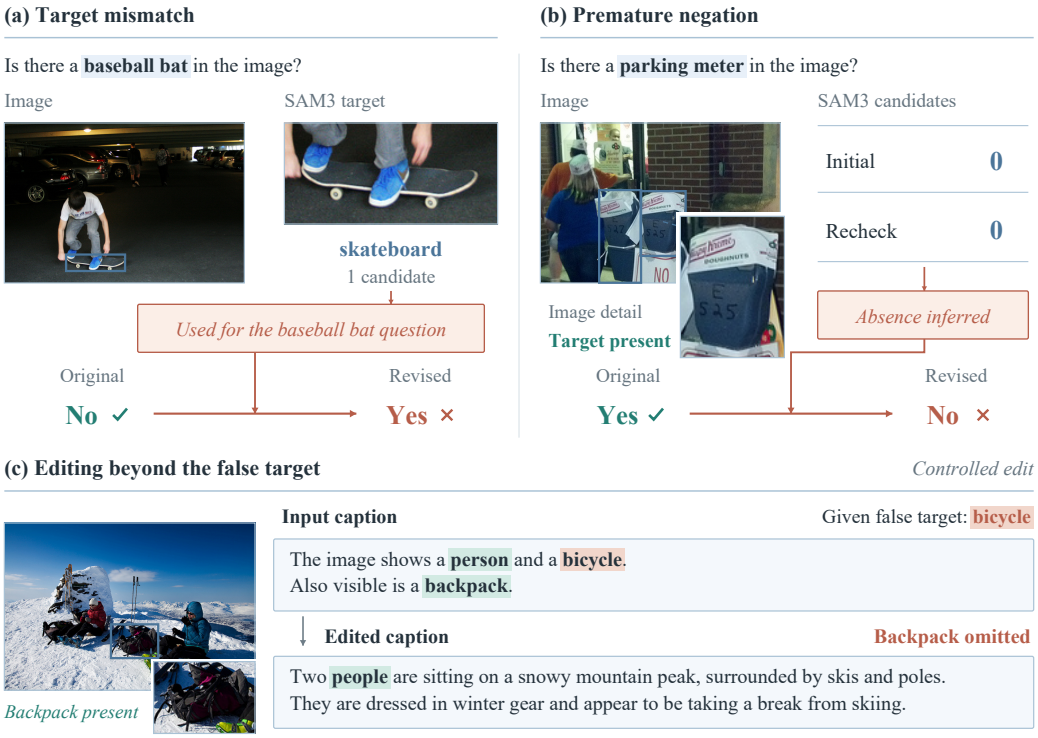

Figure 5: **Evidence–correction mismatches.** Cases from a Qwen2.5-VL and SAM 3 tool-assisted correction pipeline. (a) A skateboard detection changes a correct answer about a baseball bat. (b) Empty parking-meter observations trigger an incorrect negation. (c) The controlled input mentions a person, a bicycle, and a backpack. Given bicycle as the false target, the editor produces the displayed description of two people and omits the true backpack claim. Blue boxes are independent COCO annotations; insets magnify the source images. The controlled input and output in (c) are reproduced in full.

B.2 Target Correspondence and Empty Observations

The target-correspondence probe checks whether the tool query matches the question's target category after normalizing case, whitespace, and ordinary plurals. It retains the incoming answer when the query concerns another category or the required observation is missing from the record.

Of the 16 harmful changes prevented by target correspondence, 15 arise when observations about another target change a correct No to Yes. One concerns the distinction between *baseball bat* and *bat*. The check also forgoes three beneficial changes involving cow/giraffes, bowl/plate, and couch/chair. Its net gain is 2.71 percentage points (image-clustered 95% interval: 0.83–4.79).

Table 5 stratifies initial observations at a retention threshold of 0.50, with recorded rechecks allowed to use 0.35. Empty candidate sets and missing target records form separate strata. Harmful changes are measured at the end of the run, whose intervening decisions can also involve target mismatch.

Table 5: **Initial observation states.** Harmful changes are measured at the end of the diagnostic run.

Initial observation	Questions	Positive	Negative	Harmful changes
No retained candidates	261	21	240	19
Retained supported candidate	209	209	0	0
Target observation missing	10	10	0	1
Total	480	240	240	20

Relative to the original local rule, the empty-evidence and combined probes in Section 3 increase local decision accuracy by 1.04 percentage points (95% interval: 0.21–2.08) and 3.54 points (95% interval: 1.46–5.83), respectively. The combined gain is 17 correct decisions. The individual probes overlap on one prevented harmful change (Table 1).

The end-to-end diagnostic run yields a net gain of 14 correct answers out of 480 (image-clustered 95% interval for the accuracy change: -0.21 to 5.83 percentage points; Table 6). All 240 negative questions are answered correctly at initialization.

Table 6: **End-to-end diagnostic outcomes.** G/H count beneficial/harmful answer changes.

Label	Questions	Initial correct	Final correct	G	H
Present	240	194	223	34	5
Absent	240	240	225	0	15
Total	480	434	448	34	20

B.3 Additional-View Probes

For each of the 480 image–target pairs, separate branches start from the same initial whole-image observation. This probe retains candidates at 0.40 and assesses support at 0.50, yielding 259 empty, three weak, and 218 strong initial states. These differ from the end-to-end diagnostic thresholds used for Table 5.

Actions comprise an unchanged full-image retry, input enlargement/resampling, individual 384-pixel local tiles, and crops from observed candidates. Each image supplies two or four tiles depending on its dimensions, with one observation per tile. Input resampling retains the SAM 3 processor tensor resolution.

Table 7: **Additional-view probes.** Counts are image–target pairs, with overlap across branches. Instance matching uses IoU ≥ 0.50.

Branch	Eligible	New instance	First witness	Negative candidate
Unchanged full-image retry	480	0	0	0
Input resampling	480	4	2	0
Tile 00	480	9	3	0
Tile 01	480	9	3	0
Tile 02	384	4	1	0
Tile 03	384	5	2	1
Initial-candidate crop	221	4	0	0

A negative candidate is returned for a target explicitly annotated as absent. At IoU 0.50, the union across branches contains 18 pairs with a new matched instance and five with a first annotation-matched witness. A first witness can follow an earlier candidate that failed localization. At IoU 0.40 and 0.60, the corresponding new-instance/first-witness counts are 14/4 and 22/6, applying each threshold to both initial and additional-view matching. There are 179 pairs eligible for every action; Table 7 uses action-specific eligibility denominators.

C Experimental Details and Additional Results

C.1 Evaluation Protocols

POPE. MS-COCO, A-OKVQA (Schwenk et al., 2022), and GQA (Hudson & Manning, 2019) each contribute three subsets of 3,000 questions, one per negative-sampling strategy, with balanced positive and negative labels. The 27,000 questions span 1,489 image groups. The source-averaged scores in Section 5 average the three subset metrics within each source and then across sources. Pooled scores instead combine the confusion matrices before computing each metric.

MME. Each of the four tasks contains 30 positive/negative question pairs, giving 240 questions from 110 unique images. A task score sums question accuracy and pair accuracy, both in percent. Pair accuracy requires both answers to be correct, giving maxima of 200 points per task and 800 overall.

Captioning. We randomly sample 500 MS-COCO validation images without replacement. Both backbones use the same images over three repeats, yielding 1,500 captions per backbone. Each initial caption is paired with its corrected output.

C.2 Calibration and Threshold Sensitivity

Existence support. Support-threshold selection balances useful corrections against new errors. We maximize net correction on 2,682 calibration questions over $\{0.50, 0.55, \ldots, 0.90\}$, holding the acquired observations fixed. Table 8 also includes subsequent Qwen checks at 0.40 and 0.45. At the selected Qwen and LLaVA thresholds (0.50 and 0.70), beneficial/harmful changes are 226/56 and 428/145. Qwen groups calibration questions by image, with split sizes given in Section 5. Its acquisition threshold remains 0.90, followed by a single final adjudication at 0.50.

Table 8: **Support-threshold sensitivity.** Selected values are bold. Net is $G - H$.

	Qwen			LLaVA		
Threshold	Acc.	F1	Net	Acc.	F1	Net
0.40	89.63	89.31	164	—	—	—
0.45	89.75	89.36	167	—	—	—
0.50	**89.86**	**89.39**	**170**	89.63	89.28	277
0.55	89.82	89.31	169	89.52	89.15	274
0.60	89.08	88.40	149	89.56	89.19	275
0.65	88.89	88.11	144	89.78	89.40	281
0.70	88.29	87.37	128	**89.86**	**89.47**	**283**
0.75	87.58	86.48	109	89.75	89.34	280
0.80	87.02	85.75	94	89.19	88.69	265
0.85	86.06	84.51	68	88.52	87.90	247
0.90	85.50	83.74	53	88.29	87.61	241

Count qualification. Count qualification determines whether a conclusion supported by the candidate count can replace the initial answer. MME uses the score path $\{0.25, 0.30, 0.35, 0.40\}$ with primary candidate threshold 0.40. We scan qualification thresholds over $\{0.60, 0.65, \ldots, 0.90\}$ on the same 240 questions used for evaluation (Table 9). The qualification test applies after candidate counting, and a sufficient lower-bound witness can bypass it. Selection maximizes total MME score, breaking ties by fewer harmful changes, fewer total changes, and then the highest threshold. This selects 0.80 for both backbones. Existence thresholds come from POPE, and LLaVA's color policy retains the original answer.

Table 9: **MME count-threshold sensitivity.** The selected threshold is bold; other task scores remain fixed.

	Qwen		LLaVA	
Threshold	Count	Total	Count	Total
0.60	175.00	735.00	178.33	693.33
0.65	180.00	740.00	180.00	695.00
0.70	180.00	740.00	180.00	695.00
0.75	180.00	740.00	180.00	695.00
0.80	**180.00**	**740.00**	**180.00**	**695.00**
0.85	175.00	735.00	170.00	685.00
0.90	170.00	730.00	163.33	678.33

C.3 Detailed Benchmark Results

Tables 10 and 11 separate POPE performance by image source and negative-sampling strategy. Base denotes the unmodified backbone evaluated under the standard benchmark protocol. For each backbone, the held-out rows exclude 148 calibration images, leaving 8,136 MS-COCO, 8,046 A-OKVQA, and 8,136 GQA questions across the three sampling strategies.

Table 10: **POPE performance with Qwen2.5-VL-7B (%).** Best full-benchmark results are bolded; held-out rows exclude calibration images.

Dataset	Method	Random		Popular		Adversarial		Average	
		Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑
MS-COCO	*Base* (Bai et al., 2025)	84.30	81.42	83.60	80.75	83.23	80.41	83.71	80.86
	VCD (Leng et al., 2024)	84.70	81.90	84.00	81.35	83.67	80.92	84.12	81.39
	M3ID (Favero et al., 2024)	84.57	81.61	84.07	81.21	83.60	80.62	84.08	81.14
	SID (Huo et al., 2025)	86.13	83.92	85.13	83.25	84.47	82.35	85.24	83.17
	AvisC (Woo et al., 2025)	87.33	85.70	86.53	84.94	85.40	83.87	86.42	84.84
	DiVE (Li et al., 2026)	88.70	87.47	87.77	86.57	86.50	85.38	87.66	86.47
	EViCA	**90.53**	**89.68**	**89.23**	**88.43**	**88.20**	**87.46**	**89.32**	**88.52**
	EViCA (held-out)	90.71	89.90	89.38	88.62	88.31	87.61	89.47	88.71
A-OKVQA	*Base* (Bai et al., 2025)	88.00	86.81	86.03	84.98	80.20	79.96	84.74	83.92
	VCD (Leng et al., 2024)	88.57	87.30	86.40	85.25	81.73	81.24	85.57	84.60
	M3ID (Favero et al., 2024)	88.30	87.97	86.20	86.70	80.43	80.52	84.98	85.06
	SID (Huo et al., 2025)	87.23	86.27	86.53	85.40	82.00	81.26	85.25	84.31
	AvisC (Woo et al., 2025)	89.40	88.75	87.33	86.88	81.17	81.93	85.97	85.85
	DiVE (Li et al., 2026)	90.93	90.64	88.33	88.27	81.37	82.49	86.88	87.13
	EViCA	**93.43**	**93.20**	**92.43**	**92.24**	**87.40**	**87.71**	**91.09**	**91.05**
	EViCA (held-out)	93.21	92.96	92.24	92.03	86.95	87.28	90.80	90.76
GQA	*Base* (Bai et al., 2025)	87.93	86.88	81.23	80.99	80.83	80.66	83.33	82.84
	VCD (Leng et al., 2024)	88.03	86.64	84.63	83.53	82.03	81.24	84.90	83.80
	M3ID (Favero et al., 2024)	88.47	86.92	82.70	81.64	82.27	81.16	84.48	83.24
	SID (Huo et al., 2025)	88.27	87.20	83.27	82.47	81.90	80.94	84.48	83.54
	AvisC (Woo et al., 2025)	88.37	87.42	84.03	83.53	81.63	81.50	84.70	84.15
	DiVE (Li et al., 2026)	89.97	89.43	84.60	84.64	82.10	82.58	85.56	85.55
	EViCA	**92.30**	**91.95**	**91.13**	**90.84**	**87.40**	**87.46**	**90.28**	**90.08**
	EViCA (held-out)	92.37	92.05	91.45	91.17	87.83	87.89	90.55	90.37

Table 11: **POPE performance with LLaVA-v1.5-7B (%).** Best full-benchmark results are bolded; held-out rows exclude calibration images.

Dataset	Method	Random		Popular		Adversarial		Average	
		Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑
MS-COCO	*Base* (Liu et al., 2024)	87.09	86.14	85.13	84.28	79.20	81.21	83.81	83.88
	VCD (Leng et al., 2024)	87.20	87.34	83.47	84.30	79.27	80.23	83.31	83.96
	M3ID (Favero et al., 2024)	86.57	86.43	84.27	84.38	79.13	81.52	83.32	84.11
	SID (Huo et al., 2025)	89.10	89.21	85.97	85.29	80.33	81.24	85.13	85.25
	ONLY (Wan et al., 2025)	89.73	89.15	85.87	86.25	79.37	80.57	84.99	85.32
	AvisC (Woo et al., 2025)	86.63	84.85	85.73	83.68	82.03	81.64	84.80	83.39
	DiVE (Li et al., 2026)	86.83	85.13	85.93	84.28	83.83	82.34	85.53	83.92
	EViCA	**90.67**	**89.84**	**89.63**	**88.84**	**88.30**	**87.58**	**89.53**	**88.76**
	EViCA (held-out)	90.78	89.99	89.75	88.99	88.46	87.78	89.66	88.92
A-OKVQA	*Base* (Liu et al., 2024)	86.30	87.49	80.33	82.34	68.33	75.75	78.32	81.86
	VCD (Leng et al., 2024)	84.17	86.54	78.43	81.32	68.53	75.32	77.04	81.06
	M3ID (Favero et al., 2024)	84.53	86.27	77.80	81.21	69.30	76.21	77.21	81.23
	SID (Huo et al., 2025)	87.27	88.43	81.27	82.50	71.23	77.49	79.92	82.81
	ONLY (Wan et al., 2025)	86.13	87.18	79.07	81.05	69.03	76.94	78.08	81.72
	AvisC (Woo et al., 2025)	85.23	86.33	82.87	84.37	73.37	77.05	80.49	82.58
	DiVE (Li et al., 2026)	90.63	90.23	87.43	87.31	79.87	81.11	85.98	86.22
	EViCA	**93.53**	**93.35**	**92.63**	**92.49**	**87.07**	**87.52**	**91.08**	**91.12**
	EViCA (held-out)	93.40	93.20	92.54	92.38	86.73	87.20	90.89	90.93
GQA	*Base* (Liu et al., 2024)	85.43	86.75	79.57	81.90	68.70	74.92	77.90	81.19
	VCD (Leng et al., 2024)	84.17	85.75	74.10	76.43	69.13	75.24	75.80	79.14
	M3ID (Favero et al., 2024)	84.57	86.20	74.17	78.35	69.23	76.37	75.99	80.31
	SID (Huo et al., 2025)	85.37	88.15	74.97	80.12	69.57	76.30	76.64	81.52
	ONLY (Wan et al., 2025)	85.97	88.10	74.27	78.60	69.57	75.87	76.60	80.86
	AvisC (Woo et al., 2025)	87.43	88.25	79.43	82.53	74.67	78.13	80.51	82.97
	DiVE (Li et al., 2026)	88.83	88.32	85.00	84.92	81.03	81.66	84.95	84.97
	EViCA	**92.07**	**91.72**	**90.93**	**90.65**	**86.90**	**87.03**	**89.97**	**89.80**
	EViCA (held-out)	92.04	91.71	91.11	90.84	87.28	87.39	90.14	89.98

The task breakdown in Table 12 shows where the MME totals improve and where individual scores remain below the comparison methods.

Table 12: **Detailed MME-Hallucination results.** Best scores within each backbone are bolded.

Method	Object-level		Attribute-level		Total
	Existence	Count	Position	Color	Score↑
Qwen2.5-VL-7B					
Base (Bai et al., 2025)	195.00	165.00	140.00	190.00	690.00
VCD (Leng et al., 2024)	195.00	165.00	145.00	190.00	695.00
M3ID (Favero et al., 2024)	195.00	165.00	150.00	190.00	700.00
SID (Huo et al., 2025)	195.00	153.30	**165.00**	190.00	703.30
AvisC (Woo et al., 2025)	195.00	153.30	**165.00**	**195.00**	708.30
DiVE (Li et al., 2026)	195.00	165.00	**165.00**	**195.00**	720.00
EViCA	**200.00**	**180.00**	**165.00**	**195.00**	**740.00**

Table 12 (continued)

Method	Object-level		Attribute-level		Total
	Existence	Count	Position	Color	Score↑
LLaVA-v1.5-7B					
Base (Liu et al., 2024)	190.00	155.00	128.30	**170.00**	643.30
VCD (Leng et al., 2024)	190.00	155.00	133.30	**170.00**	648.30
M3ID (Favero et al., 2024)	**195.00**	160.00	133.30	165.00	653.30
SID (Huo et al., 2025)	190.00	158.30	136.70	165.00	650.00
ONLY (Wan et al., 2025)	**195.00**	160.00	133.30	165.00	653.30
AvisC (Woo et al., 2025)	**195.00**	156.70	143.30	**170.00**	665.00
DiVE (Li et al., 2026)	**195.00**	163.30	143.30	**170.00**	671.70
EViCA	**195.00**	**180.00**	**165.00**	155.00	**695.00**

C.4 Additional POPE Metrics and Ablations

Additional metrics. Precision and recall in Table 13 characterize the errors that remain after correction. These metrics pool all 27,000 POPE questions. Precision exceeds recall for both backbones, and the fractions of yes answers are 46.48% for Qwen and 46.85% for LLaVA. On this balanced benchmark, the remaining errors therefore include more missed positives than false positives.

Table 13: **Additional POPE metrics for EViCA (%).** Results cover all 27,000 questions; Net is $G - H$.

Backbone	Acc.	F1	Precision	Recall	G	H	Net
Qwen	90.23	89.87	93.27	86.71	2,419	747	1,672
LLaVA	90.19	89.87	92.89	87.04	4,067	1,297	2,770

Variant definitions. The ablations share initial Qwen responses, support thresholds, and answer parsers. *Initial only* uses the initial recorded observation with Full's adjudicator, leaving further acquisition requirements unmet. *Score only* uses all observations acquired by Full but bases its decision on the support score. *Fixed schedule* retains the action types, budgets, and final adjudicator, but freezes a first-in, first-out request queue from the initial image geometry and candidates. Full regenerates requests from the current unmet evidence conditions at each round; Fixed schedule advances through its frozen queue. Both stop when no unmet conditions remain or the budget is exhausted.

Uncertainty estimates. To account for questions sharing an image, we resample image groups with replacement 10,000 times. Each draw retains all questions for the sampled images and is shared across variants. We recompute paired metric differences for every draw and report their 2.5th and 97.5th percentiles (Table 14).

Table 14: **Paired ablation differences.** Full minus each variant, in percentage points, with 95% intervals.

Comparison	Metric	Difference	95% interval
Full − Initial only	Accuracy	$+0.6908$	$[+0.3406, +1.0288]$
	F1	$+0.3496$	$[-0.0195, +0.7035]$
Full − Score only	Accuracy	$+0.4976$	$[+0.2905, +0.7098]$
	F1	$+0.6462$	$[+0.4131, +0.8842]$
Full − Fixed schedule	Accuracy	-0.0288	$[-0.1400, +0.0742]$
	F1	-0.0626	$[-0.1842, +0.0470]$

Weak evidence and adjudication states. Full's weak-evidence branch contains 591 Unknown questions with nonempty evidence and valid support below 0.50. The 287 original yes answers, spanning 129 images, account for all answer differences between Full and Score only. The remaining 304 original no answers agree under Full and Score only. Table 16 reports adjudication states for

Table 15: **Correction direction on Qwen POPE.** Results cover 24,318 questions; Net is $G - H$.

Variant	Direction	G	H	Net
Full	No → Yes	1,678	336	1,342
	Yes → No	515	355	160
Initial only	No → Yes	1,651	317	1,334
	Yes → No	0	0	0
Score only	No → Yes	1,678	336	1,342
	Yes → No	598	559	39
Fixed schedule	No → Yes	1,720	371	1,349
	Yes → No	515	355	160

Full, Initial only, and Fixed schedule; Score only makes a binary decision without assigning these four states.

Table 16: **Adjudication states on Qwen POPE.** Calibration images are excluded.

Variant	Supported	Refuted	Unknown	Conflict
Full	11,034	12,663	621	0
Initial only	10,946	0	13,372	0
Fixed schedule	11,223	12,663	432	0

Performance and observation requests. On the held-out set, Fixed schedule obtains seven additional net corrections, with final answers differing from Full on 77 questions. This small gain accompanies a higher number of logical observation requests. Requests include reused observations, while latency is measured separately in Table 20. Table 17 reports the ablations on the complete 27,000-question benchmark, including calibration questions.

Table 17: **Full-benchmark Qwen ablations.** All 27,000 questions are included.

Variant	Acc.	F1	G	H	Net	Requests
Initial only	89.52	89.50	1,825	344	1,481	27,000
Score only	89.77	89.26	2,513	966	1,547	57,028
Fixed schedule	**90.24**	**89.92**	2,461	785	1,676	83,047
EViCA (Full)	90.23	89.87	2,419	747	1,672	57,028

C.5 Caption Quality and Content Retention

For each repeat, CHAIR_S is the fraction of captions containing a hallucinated COCO object mention; CHAIR_I divides hallucinated object mentions by all generated object mentions. Object recall divides the sum of correctly mentioned distinct ground-truth categories per image by the summed ground-truth category count. We use official CHAIR resources and the standalone implementation's recall metric, averaging rates across three repeats.

Tables 18 and 19 report scores across repeats and caption editing statistics. All repeats use the same 500 images, giving 1,500 captions per backbone. Mean words uses whitespace-delimited words. Token ratio is the mean, over captions, of the edited token count divided by the original token count; tokens are produced by NLTK's TreebankWordTokenizer.

Table 18: **Caption quality across three repeats (%).**

Backbone	Repeat	$CHAIR_S$	$CHAIR_I$	Recall
Qwen	1	26.80	6.927	65.96
	2	24.60	6.870	65.77
	3	24.40	6.426	65.89
LLaVA	1	37.60	9.429	76.16
	2	34.20	8.419	76.22
	3	35.80	8.825	75.09

Table 19: **Caption length and editing statistics.**

Backbone	Mean words	Token ratio	Changed	Deleted mentions
Qwen	144.504	91.71%	989	2,195
LLaVA	79.452	87.90%	952	2,129

C.6 Inference Latency

We decompose inference latency into generation, claim processing, evidence processing, and correction. The profile covers 500 images with Qwen and SAM 3 resident, starting after model loading and warm-up. Let g_i, c_i, e_i, r_i denote generation, claim processing, evidence processing, and correction times for image i, in seconds, and let n_i be its number of initially generated tokens. We report

$$L = \frac{1000 \sum_i (g_i + c_i + e_i + r_i)}{\sum_i n_i} \quad \text{ms per initially generated token.} \tag{7}$$

The denominator includes special tokens, excludes input tokens, and remains fixed during caption deletion. Combined latency sums the four separately measured stages.

Table 20: **Stage-level latency of EViCA.** Results cover 500 images.

Stage	Latency (ms/token)
Generation	30.0081
Claim processing	0.0639
Evidence and adjudication	47.6070
Correction	0.0533
Combined	77.7322

Table 21 reports GPU memory and average latency for the compared methods.

Table 21: **Computational cost on Qwen2.5-VL-7B.** GPU memory and average latency.

Method	GPU memory (MB)	Avg. latency (ms/token)
Base (Bai et al., 2025)	17,568	33.4
VCD (Leng et al., 2024)	18,513	65.9
SID (Huo et al., 2025)	18,118	68.4
M3ID (Favero et al., 2024)	18,309	65.1
AvisC (Woo et al., 2025)	17,999	67.1
DiVE (Li et al., 2026)	18,078	42.6
EViCA	21,806	77.7