NOT BLIND BUT SILENCED: PRESERVING VISUAL EVIDENCE THROUGH COUNTER-CONTEXT VALIDATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) can produce incorrect answers even when intermediate representations robustly favor the factual hypothesis. We characterize this failure through layer-wise evidence competition and identify *strict silencing*, in which established intermediate factual dominance is overturned before the final prediction. To preserve the influence of useful visual evidence, we propose **Adaptive Counter-context Evidence (ACE)**, a training-free decoding framework that follows *Challenge*, *Validate*, *Locate*, and *Preserve*. ACE constructs a controlled counter-context to estimate contribution reliability, locates an evidence anchor, and selectively amplifies original-image contributions throughout the decoder suffix. The counter-context is discarded before answer generation, while reliability gates remain fixed and current contributions evolve with each decoding step. Across four MLLM backbones, ACE improves POPE-Adversarial accuracy by 6.37–8.97 percentage points and reduces CHAIR_s by 3.9–5.0 points, while maintaining MM-Vet performance within 0.1 point of greedy decoding. Functional diagnostics demonstrate stronger evidence identification, while ACE achieves four times as many strict-silencing rescues as uniform-suffix preservation on the fixed Qwen3 evaluation cohort.

1 INTRODUCTION

Multimodal large language models (MLLMs) have achieved strong visual understanding capabilities (Alayrac et al., 2022; Liu et al., 2023; Li et al., 2023a; Dai et al., 2023; Zhu et al., 2024), yet object hallucinations and broader perceptual failures persist (Rohrbach et al., 2018; Li et al., 2023c; Guan et al., 2024; Wang et al., 2023; Huang et al., 2024b; Tong et al., 2024). Existing decoding and internal interventions mitigate these failures through altered views, attention correction, and representation adjustment (Leng et al., 2024; Huang et al., 2024a; Wang et al., 2025a; Zhang et al., 2026; Lin et al., 2026). Recent studies further show that useful intermediate visual evidence can coexist with an incorrect final prediction (Wang et al., 2025a; Kan et al., 2026; Li et al., 2026; Guo et al., 2026). This raises our first question: *how can an intermediate factual advantage be lost before the final decision?*

We characterize this discrepancy through competition between factual and competing hypotheses across decoder depth. A robust intermediate factual advantage may be overturned before terminal prediction, producing what we call *strict silencing* when the raw answer is incorrect and the terminal layer robustly favors the competitor. The resulting margin decomposition quantifies the downstream shift required to overturn an established intermediate factual advantage, providing a precise characterization of terminal takeover.

Our second question is *which contributions should be preserved?* Attention alone need not indicate factual support (Jha & Kim, 2026; Wang et al., 2026a), while high-norm visual representations may participate in useful computation (Darcet et al., 2024; Jiang et al., 2025). We therefore construct a controlled counter-context probe and estimate contribution reliability from decision coupling, counter-context stability, and visual specificity. These label-free proxies guide selection; single-unit ablation factual effect (AFE) separately assesses their functional relevance.

Our third question is *how to preserve selected evidence through subsequent computation?* We propose **Adaptive Counter-context Evidence (ACE)**, a training-free framework comprising *Challenge*, *Validate*, *Locate*, and *Preserve*. ACE locates an evidence anchor, discards the counter-context before the first answer token, and selectively amplifies current original-image contributions throughout the

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.

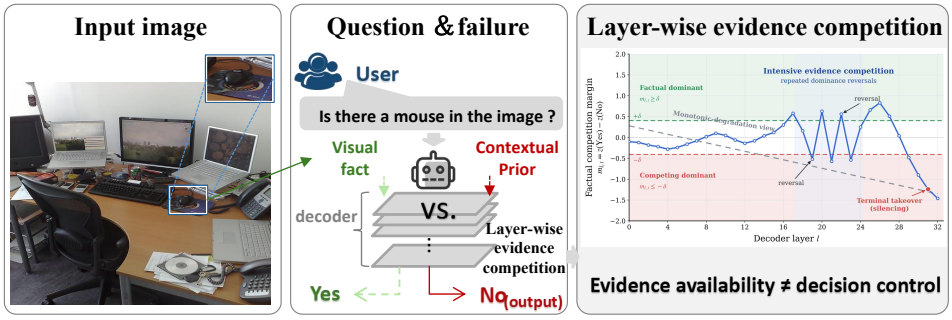


Figure 1: **Conceptual illustration of evidence competition and silencing.** Although the image supports the factual answer "Yes," factual and competing hypotheses can repeatedly exchange dominance across decoder depth. A robust intermediate factual advantage may still be overturned before terminal prediction, illustrating that *evidence availability does not guarantee decision control*.

decoder suffix. Reliability gates remain fixed across generation steps, whereas the contributions themselves are continuously recomputed.

Our contributions are threefold: **(1) Silencing characterization.** We formalize robust evidence competition and identify strict silencing as a measurable failure of sustained factual dominance. **(2) Counter-context evidence identification.** We develop label-free reliability estimation and anchor localization, with functional ablation providing independent evaluation of candidate visual contributions. **(3) Sustained selective preservation.** We introduce original-image decoder-suffix preservation and demonstrate its effectiveness through strict-silencing recovery and controlled reliability-allocation experiments.

2 Related Work

Intermediate evidence and hallucination dynamics. DeCo, RAR, DiVE, and FADE investigate intermediate-layer correction and visual-evidence dynamics (Wang et al., 2025a; Kan et al., 2026; Li et al., 2026; Guo et al., 2026). Building on these findings, ACE characterizes robust dominance reversals and targets sustained visual-evidence influence. Related benchmarks assess multimodal understanding and reasoning (Fu et al., 2025; Liu et al., 2024c; Li et al., 2024).

Functional evidence and internal intervention. Existing approaches intervene in attention pathways (Ji et al., 2026; Chang et al., 2026), internal representations (Wang et al., 2026b; Lin et al., 2026; Dang et al., 2026), or prefill caches (Zhang et al., 2026). Since attention alone does not establish factual support (Jha & Kim, 2026; Wang et al., 2026a), ACE uses controlled counter-context validation to identify candidate visual contributions and guide anchor-based, reliability-weighted preservation.

Altered views and evidence preservation. Altered-view decoding (Leng et al., 2024; Liang et al., 2026; Chen et al., 2024), visual attention steering (Huang et al., 2024a; Liu et al., 2024b; 2026), and alignment-based correction (Liu et al., 2024a; Yu et al., 2024a; Sun et al., 2024; Yin et al., 2024; Yue et al., 2024) address related hallucination failures. Unlike altered-view decoding, ACE discards the counter-context before generation and intervenes exclusively on the original-image decoder trajectory.

Readouts, interventions, and attribution. Layer-wise readouts and activation interventions support latent-prediction analysis and representation control (Chuang et al., 2024; Belrose et al., 2023; Li et al., 2023b; Zhang et al., 2024; Rimsky et al., 2024; Meng et al., 2022). Given the limitations of attention-based attribution (Jain & Wallace, 2019; Abnar & Zuidema, 2020; Chefer et al., 2021), ACE distinguishes label-dependent diagnostics, label-free runtime selection, and post hoc functional ablation.

3 MECHANISTIC FORMULATION

We characterize visual hallucination through layer-wise competition between factual and competing hypotheses. Tracking their relative dominance reveals when a robust intermediate factual advantage is overturned before terminal decoding. This perspective establishes the diagnostic foundation for identifying and preserving recoverable visual evidence.

3.1 LAYER-WISE EVIDENCE COMPETITION AND SILENCING

At generation position t, let y_f denote the factual hypothesis consistent with the visual input and y_c its mutually exclusive competitor. Let $\mathcal{T}(y)$ denote the lexical realizations of hypothesis y. The two lexical sets are fixed, nonempty, and disjoint; for binary answer diagnostics they contain distinct first-token realizations. For decoder hidden state $h_{\ell,t}$, we define

$$\mathbf{z}_{\ell,t} = W_U \,\mathrm{Norm}(h_{\ell,t}), \qquad s_{\ell,t}(y) = \log \sum_{u \in \mathcal{T}(y)} \exp z_{\ell,t}(u), \qquad m_{\ell,t} = s_{\ell,t}(y_f) - s_{\ell,t}(y_c). \tag{1}$$

Positive $m_{\ell,t}$ favors the factual hypothesis, while negative $m_{\ell,t}$ favors its competitor. Given a tolerance $\delta > 0$, we call a layer factual-dominant when $m_{\ell,t} \geq \delta$, competing-dominant when $m_{\ell,t} \leq -\delta$, and weak-band otherwise. After removing weak-band layers, a *dominance reversal* occurs whenever two consecutive robust states have opposite signs.

We identify *silencing* through a stricter condition that requires established intermediate factual dominance to be overturned by robust competing dominance at the terminal layer:

$$\mathcal{S}_{\mathrm{sil}} = \left\{ \max_{\ell < L} m_{\ell,t} \geq \delta \;\wedge\; m_{L,t} \leq -\delta \right\}, \tag{2}$$

The empirical strict-silencing cohort additionally requires an incorrect prediction by the unmodified model: $\mathcal{E}^{\mathrm{raw}}_{\mathrm{sil}} = \mathcal{E}^{\mathrm{raw}}_{\mathrm{err}} \cap \mathcal{S}_{\mathrm{sil}}$. This definition combines an observed prediction error with a robust intermediate-to-terminal dominance transition. We denote the strongest pre-terminal factual state by $\ell^\star = \arg\max_{\ell < L} m_{\ell,t}$.

3.2 RECOVERABLE EVIDENCE AND TERMINAL TAKEOVER

Intermediate factual dominance identifies where a factual hypothesis becomes accessible. We further characterize how its advantage evolves from the strongest pre-terminal factual state $\ell^\star$ to the terminal prediction. Defining $\Delta m_{k,t} = m_{k+1,t} - m_{k,t}$, we obtain

$$m_{L,t} = m_{\ell^\star,t} + \sum_{k=\ell^\star}^{L-1} \Delta m_{k,t}. \tag{3}$$

Equation 3 characterizes *terminal takeover*: strict silencing requires the cumulative downstream margin change to reach $-(m_{\ell^\star,t} + \delta)$ or below. This exact decomposition quantifies the shift required to overturn an established factual advantage (Appendix A.3).

This characterization defines the target failure condition and motivates ACE's inference-time objective: identifying candidate visual evidence and sustaining its influence throughout downstream decoder computation. ACE estimates reliability without diagnostic labels, while independent functional ablations and prediction-recovery experiments evaluate evidence identification and the resulting intervention.

4 METHOD

We introduce **Adaptive Counter-context Evidence (ACE)**, a training-free inference-time framework comprising four stages: *Challenge*, *Validate*, *Locate*, and *Preserve*.

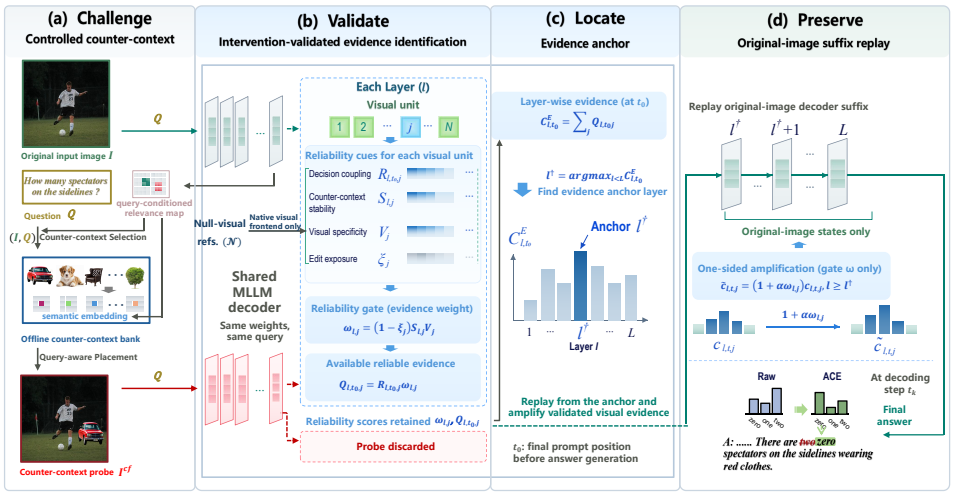


Figure 2: **ACE framework. Challenge** constructs a controlled counter-context probe; **Validate** estimates visual-unit reliability at the final prompt position t_0; **Locate** identifies the evidence anchor $\ell^\dagger$; and **Preserve** discards the probe and replays the original-image decoder suffix with sustained selective amplification.

ACE constructs a controlled counter-context probe, estimates label-free contribution reliability, and locates an evidence anchor. It then discards the probe and replays the original-image decoder suffix with selective amplification. Validation evaluates contribution reliability using the defined proxy criteria, while answer generation operates exclusively on the original-image branch.

4.1 Challenge: Controlled Counter-Context Construction

ACE begins with a standard prefill on the original image-question pair (I, q). Let t_0 denote the final prompt position before answer generation and $j \in \mathcal{I}_V$ a decoder-accessible visual-conditioned unit. No answer token is selected at this stage.

For attention head h, let $a^{(h)}_{\ell,t,j}$ and $v^{(h)}_{\ell,j}$ denote the attention weight and value associated with unit j. Unit indices are aligned across the original, counter-context, and null views under the same preprocessing geometry. We define its projected contribution to the attention readout as

$$c_{\ell,t,j} = \sum_{h=1}^{H} a^{(h)}_{\ell,t,j} W^{(h)}_{O,\ell} v^{(h)}_{\ell,j}. \tag{4}$$

where $W^{(h)}_{O,\ell}$ denotes the output-projection submatrix corresponding to attention head h. Unlike attention probability alone, $c_{\ell,t,j}$ captures both attention allocation and value content. Using original-image contributions at t_0, ACE constructs a query-conditioned relevance map and places a semantically contrasting asset from the fixed offline Counter-Context Intervention Bank (CCIB) in a low-relevance region. Candidates matching the queried concept are excluded.

Probe acceptance jointly enforces relevance-weighted feature preservation outside the edited support and bounded contextual displacement: $P_{\text{pres}} \geq \tau_{\text{pres}}$ and $D_C^{\min} \leq D_C \leq D_C^{\max}$. These criteria establish a controlled challenge that changes the surrounding context while limiting changes to estimated relevant features outside the edit.

If no candidate is admissible, ACE returns to standard decoding. The construction and acceptance protocols are detailed in Appendices B.3 and B.4.

Candidate validity uses a separate frozen visual encoder and requires no candidate-level language-decoder prefill. Only the selected admissible probe enters the MLLM reliability pass; the three null references use the native visual frontend only (Appendix B).

4.2 VALIDATE AND LOCATE: EVIDENCE IDENTIFICATION

Given an accepted counter-context, ACE evaluates each visual-conditioned unit using *decision coupling*, *counter-context stability*, and *visual specificity*. Decision coupling is measured at t_0, stability from paired original and modified prefills, and visual specificity against fixed null-visual references.

Decision coupling is a magnitude-based participation proxy,

$$R_{\ell,t_0,j} = \frac{\|c_{\ell,t_0,j}\|_2}{\sum_{i\in\mathcal{I}_{\text{ctx}}} \|c_{\ell,t_0,i}\|_2 + \epsilon}. \tag{5}$$

This label-free score measures each unit's relative contribution magnitude and provides the participation component of ACE's evidence-ranking criterion. Counter-context stability is

$$S_{\ell,j} = \frac{1 + \cos\left(v_{\ell,j}^{\text{cat,raw}}, v_{\ell,j}^{\text{cat,cf}}\right)}{2}, \qquad v_{\ell,j}^{\text{cat}} = \text{Concat}\left(v_{\ell,j}^{(1)}, \ldots, v_{\ell,j}^{(H)}\right). \tag{6}$$

Visual specificity estimates dependence on the input image through comparison with null-visual references,

$$V_j = \frac{1 - \max_{I^{\varnothing}\in\mathcal{N}} \cos\left(x_j^V(I,q), x_j^V(I^{\varnothing},q)\right)}{2}, \tag{7}$$

where $\mathcal{N}$ is a fixed set of null-visual references and x_j^V is the decoder-facing visual embedding before the language decoder, after the native visual projection and packing operations. Appendix B.5 fixes this feature interface. Let $\xi_j \in [0,1]$ denote the direct exposure of unit j to the edited support. ACE defines a reliability gate and a candidate-evidence score as

$$\omega_{\ell,j} = (1-\xi_j) S_{\ell,j} V_j, \qquad Q_{\ell,t_0,j} = R_{\ell,t_0,j}\omega_{\ell,j}. \tag{8}$$

The candidate-evidence score $Q_{\ell,t_0,j}$ integrates contribution magnitude, counter-context stability, visual specificity, and direct edit exposure to identify candidate reliable visual contributions.

ACE uses these label-free estimates for runtime evidence selection. Their factual relevance is evaluated independently through single-unit ablation factual effect (AFE), which measures the terminal-margin change induced by deleting an individual contribution (Appendix C.1). Neither AFE nor factual labels enter runtime reliability estimation or anchor localization.

ACE then aggregates candidate-evidence scores and selects a label-free anchor (ties favor the shallower layer),

$$C_\ell^E = \sum_{j\in\mathcal{I}_V} Q_{\ell,t_0,j}, \qquad \ell^\dagger = \arg\max_{\ell<L} C_\ell^E. \tag{9}$$

If the maximum evidence score is numerically negligible, ACE returns standard decoding. $Q_{\ell,t_0,j}$ is used only for anchor localization. During subsequent decoding, the current contribution $c_{\ell,t,j}$ already reflects the evolving participation of unit j, so ACE uses $\omega_{\ell,j}$ rather than $Q_{\ell,t_0,j}$ as the downstream reliability gate.

4.3 PRESERVE: TRAJECTORY-AWARE EVIDENCE PRESERVATION

After locating $\ell^\dagger$, ACE discards the counter-context branch and returns entirely to the original-image trajectory. Before selecting the first answer token, it replays the required original-image decoder suffix from the anchor and modifies each visual-conditioned contribution according to

$$\widetilde{c}_{\ell,t,j} = (1 + \alpha\omega_{\ell,j})\, c_{\ell,t,j}, \qquad \ell \geq \ell^\dagger, \tag{10}$$

where $\alpha \geq 0$ controls the preservation strength. The layer-specific gate $\omega_{\ell,j}$ is estimated once during validation, whereas $c_{\ell,t,j}$ is recomputed from the evolving original-image state at every generation step. ACE is therefore trajectory-aware without repeatedly evaluating the counter-context.

ACE employs one-sided amplification: contributions receiving higher reliability weights obtain stronger amplification, while lower-weighted contributions remain closer to their original magnitudes.

This intervention operates exclusively on the original-image decoding trajectory. Its effects propagate through subsequent decoder computation, enabling sustained influence of the selected visual evidence. Cache reuse, replay boundaries, fallback conditions, and architecture-specific mappings are detailed in the Appendix.

5 Experiments

5.1 Experimental Setup

The evaluation covers Qwen3-VL-8B, InternVL3.5-8B, Qwen2.5-VL-7B, and LLaVA-OneVision-7B configurations (Bai et al., 2025a; Wang et al., 2025b; Bai et al., 2025b; Li et al., 2025). POPE, CHAIR, and FREAK assess hallucination, and MM-Vet assesses broader multimodal capability (Li et al., 2023c; Rohrbach et al., 2018; Yin et al., 2026; Yu et al., 2024b). Main POPE results use the Adversarial split; each configuration has 3,000 paired cases with 1,500 Yes and 1,500 No labels. Random/Popular results and benchmark-specific denominators are given in the Appendix.

The main comparison includes DeCo, CausalLens, HulluEdit, and EnAR (Wang et al., 2025a; Ji et al., 2026; Lin et al., 2026; Liang et al., 2026); additional VCD, OPERA, PTI, and LTS-FS results appear in the Appendix (Leng et al., 2024; Huang et al., 2024a; Zhang et al., 2026; Dang et al., 2026). Mechanistic diagnostics concern the first-answer decision at t_0, with $\delta = \log(0.6/0.4)$ and preservation strength $\alpha = 0.35$. ACE requires no parameter update and uses a fixed offline CCIB. All subset analyses retain the identifiers and predictions of their corresponding full cohorts.

The reference execution configuration targets one NVIDIA A800 GPU, with BF16 model computation and FP32 reliability statistics. Appendix B.10 specifies the software recipe; Appendix G distinguishes fixed-shape cost profiles from benchmark-specific image preprocessing.

5.2 Main Results

Across four backbones, ACE delivers 6.37–8.97 percentage-point improvements in POPE-Adversarial accuracy, reduces CHAIR_s by 3.9–5.0 points, and improves FREAK accuracy by 2.3–2.8 points (Table 1). MM-Vet scores remain within 0.1 point of greedy decoding, demonstrating capability retention under the evaluated protocols.

Paired POPE accounting records 191, 218, 269, and 266 additional correct predictions out of 3,000 cases for the four respective backbones. These aggregate gains are complemented by the targeted strict-silencing analysis in Sec. 5.6. Benchmark-specific aggregation conventions are detailed in Appendix C.5. Example-level paired ACE–EnAR accuracy contrasts and their uncertainty estimates are reported in Appendix D.2.

5.3 Inference Efficiency

We evaluate the complete online inference cost of ACE, including candidate validation, null-visual processing, and decoder-suffix replay. For short-answer workloads, the measured slowdowns are $1.63\times$/$1.57\times$ at one token and $1.55\times$/$1.49\times$ at eight tokens on Qwen3/LLaVA.

Table 2 reports matched 64-token workloads, where ACE requires 2.93 s and 2.59 s end-to-end, corresponding to $1.34\times$ and $1.32\times$ greedy-decoding latency. Detailed stage-wise costs, memory measurements, and response-length analyses are provided in Appendices D.7 and G.

5.4 Evidence Competition and Silencing

Figure 3(b) reveals a consistent pattern across four backbones: robust dominance reversals occur in 65.0–70.0% of raw errors, compared with 30.9–36.5% of correct predictions.

Table 1: **Main results.** POPE reports Adversarial Accuracy/F1; CHAIR reports sentence/instance hallucination rates; FREAK reports Accuracy/HalluRate. Higher is better for accuracy, F1, and MM-Vet, and lower is better for hallucination rates. CHAIR instance rates use method-specific mention totals. FREAK first averages the six option permutations within each multiple-choice question; each original question has equal aggregate weight. See Appendix C.5.

Method	POPE-A Acc./F1 ↑	CHAIR S/I ↓	FREAK Acc./HR	MM-Vet ↑
Qwen3-VL-8B				
Greedy	87.93 / 87.61	22.8 / 6.3	45.5 / 41.6	67.3
DeCo	88.57 / 88.32	21.9 / 7.3	45.5 / 41.3	67.1
CausalLens	88.93 / 88.74	20.8 / 6.6	45.8 / 41.1	67.4
HulluEdit	89.33 / 89.06	19.7 / 5.8	46.5 / 40.3	67.2
EnAR	89.50 / 89.28	20.1 / 5.5	46.9 / 40.0	66.8
ACE	94.30 / 94.21	18.9 / 5.9	47.8 / 38.8	67.3
InternVL3.5-8B				
Greedy	87.03 / 86.72	24.1 / 7.2	42.8 / 44.3	83.1
DeCo	87.70 / 87.42	23.0 / 6.0	43.1 / 43.8	82.8
CausalLens	88.10 / 87.91	22.2 / 6.4	43.8 / 42.9	83.2
HulluEdit	88.47 / 88.19	21.0 / 6.2	44.1 / 42.8	83.0
EnAR	88.80 / 88.54	21.4 / 6.9	44.7 / 42.0	82.5
ACE	94.30 / 94.22	20.2 / 6.0	45.4 / 41.3	83.0
Qwen2.5-VL-7B				
Greedy	85.43 / 85.12	30.6 / 8.8	35.8 / 44.6	67.1
DeCo	86.27 / 85.96	29.4 / 6.9	36.2 / 43.7	66.8
CausalLens	87.00 / 86.71	28.0 / 7.0	36.9 / 43.5	67.0
HulluEdit	87.73 / 87.44	26.7 / 8.2	37.0 / 43.1	66.9
EnAR	88.10 / 87.82	27.2 / 7.3	37.9 / 42.3	66.4
ACE	94.40 / 94.35	25.6 / 6.5	38.6 / 41.9	67.2
LLaVA-OV-7B				
Greedy	84.17 / 83.88	28.6 / 8.0	34.3 / 47.9	57.5
DeCo	85.07 / 84.75	27.2 / 7.9	34.6 / 47.6	57.2
CausalLens	85.60 / 85.31	25.9 / 7.5	35.1 / 46.7	57.4
HulluEdit	86.20 / 85.91	24.6 / 8.6	35.6 / 46.0	57.3
EnAR	86.47 / 86.18	25.1 / 7.3	35.9 / 45.7	56.9
ACE	93.03 / 92.97	23.8 / 6.1	36.8 / 45.1	57.4

Table 2: **Inference efficiency.** Mean E2E latency and slowdown on 128 paired fixed-shape requests per profile (64 output tokens, batch size 1).

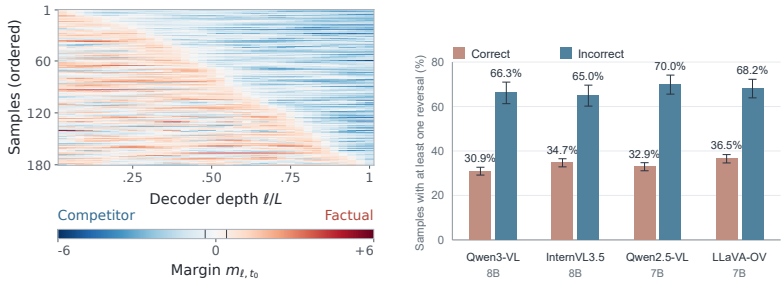
Profile	Metric	Greedy	DeCo	CausalLens	HulluEdit	EnAR	**ACE**
Qwen3	E2E (s)	2.18	2.51	2.39	2.44	4.08	**2.93**
	Slowdown	1.00×	1.15×	1.10×	1.12×	1.87×	**1.34**×
LLaVA	E2E (s)	1.96	2.23	2.12	2.18	3.61	**2.59**
	Slowdown	1.00×	1.14×	1.08×	1.11×	1.84×	**1.32**×

Among 362 Qwen3 raw errors, 180 (49.7%) satisfy strict silencing, forming the trajectories in Fig. 3(a). The other three backbones contain 187, 225, and 251 strict-silencing cases, respectively.

These observations establish a measurable failure population characterized by the loss of robust intermediate factual dominance, providing the basis for subsequent functional diagnostics and targeted recovery experiments.

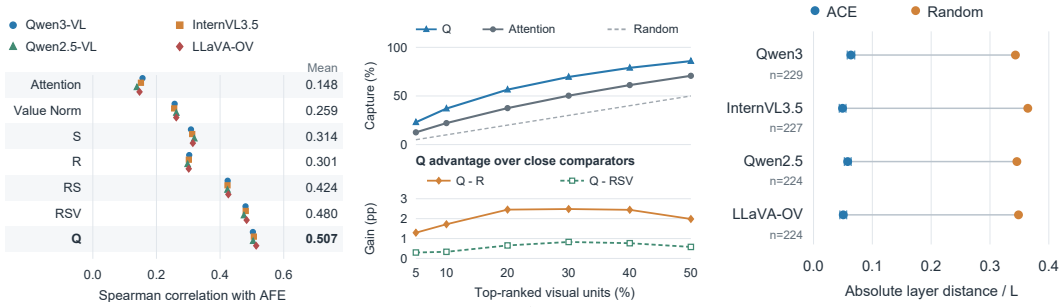
5.5 Validating Recoverable Evidence

We evaluate evidence identification through single-unit ablation factual effect (AFE), which measures the terminal factual-margin change induced by removing an individual visual contribution and replaying the downstream decoder suffix.



(a) (b)

Figure 3: **Evidence competition and silencing.** (a) The fixed 180 strict trajectories, sorted by their last preterminal layer reaching $+\delta$. (b) At least one robust reversal in each 3,000-case cohort, grouped by raw correctness; bars show 95% Wilson intervals.



(a) (b) (c)

Figure 4: **Recoverable-evidence diagnostics.** (a) Spearman alignment with AFE; color/shape identifies configuration, and the right column gives the equal-weight mean. (b) Upper: positive-AFE capture; dashed random is its exact expectation. Lower: $Q - R$ and $Q - RSV$ differences in percentage points on a separate scale. (c) Anchor-to-peak distance divided by depth, with 95% bootstrap intervals.

Figure 4 demonstrates that Q provides substantially stronger alignment with functional evidence than attention alone. Across four backbones, its mean Spearman correlation reaches 0.507 versus 0.148 for attention, while positive-AFE capture at a 10% selection budget reaches 37.0% versus 22.1%. ACE-selected anchors are also concentrated near the diagnostic factual peak: their normalized distances are 0.049–0.064, compared with 0.343–0.364 for random anchors.

Functional evidence within the target failure population. We further evaluate the fixed Qwen3 strict-silencing cohort. Among the 170 actively intervened cases, the top 10% of contributions ranked by Q achieve a mean signed AFE of 0.286, compared with 0.137 for attention, and capture 37.1% of positive AFE.

These case-aligned measurements connect ACE's reliability ranking with the functional contributions available in its target failure population. The subsequent preservation experiments evaluate how sustaining these contributions changes final predictions. Detailed within-cohort associations and statistical intervals are reported in Appendix H.1.

5.6 Ablation and Terminal-Takeover Prevention

Table 3(A) evaluates controlled counter-context construction on the same 284 valid-probe cases, using identical visual units and AFE targets. Controlled ACE achieves an AFE rank correlation of 0.503, compared with 0.407 for the strongest alternative, semantic-near construction. Alternative constructions bypass validity rejection to maintain the shared comparison cohort. Panel B evaluates preservation strategies on the same 180 Qwen3 raw strict-silencing cases, including ten fallbacks, with $\alpha = 0.35$. The comparison distinguishes one-shot intervention, uniform suffix amplification, and sustained selective preservation.

Table 3: **Mechanistic ablations for the Qwen3 configuration.** A: All rows share 284 cases admitted by the controlled-probe validity criteria and identical AFE targets. Correlation uses 282 cases after constant-score exclusions; capture uses all 284 cases. B: All 180 strict cases, including ten fallbacks; gains are averaged over the complete set.

A. Counter-context diagnostic

Construction	P_{pres}	D_C	ρ	Cap.@10%
Identity	1.000	0.000	0.377	35.7
Random	0.861	0.479	0.367	33.3
Semantic-near	0.948	0.180	0.407	36.5
Semantic-far	0.807	0.720	0.321	31.5
Controlled ACE	0.919	0.439	0.503	37.3

B. Preservation strategy

Strategy	Rescued	Rescue	$\overline{\Delta m}_L$
Final-layer	3	1.7%	0.52
Anchor only	9	5.0%	0.71
Uniform suffix	13	7.2%	0.82
Anchor-frozen gate	49	27.2%	1.34
ACE	52	28.9%	1.42

Table 3(B) demonstrates the contribution of sustained selective preservation. ACE achieves 52/180 strict rescues, four times the 13/180 obtained with uniform-suffix preservation and substantially above the 9/180 obtained with anchor-only intervention. The anchor-frozen variant also achieves substantial recovery (49/180), confirming the importance of maintaining selective intervention throughout the decoder suffix. Layer-specific reliability further increases the mean terminal factual margin by 0.084 (95% interval $[0.061, 0.107]$). The corresponding binary strict-rescue difference is smaller and is reported with its paired significance test in Appendix H.2.

ACE consistently recovers strict-silencing cases across all four backbones, achieving 52/180, 75/187, 94/225, and 80/251 strict rescues, respectively (Table 13). For Qwen3, the 191 net corrections include 52 strict rescues, 19 additional corrections within the strict-silencing cohort, and 120 corrections outside that cohort. This decomposition establishes both targeted recovery under the defined strict criterion and broader gains in benchmark prediction accuracy. Complete paired transition accounting is provided in Appendix D.8.

To isolate reliability allocation from amplification strength, we evaluate amplitude-matched controls with fixed probes, anchors, and fallback decisions. At the $\alpha_0 = 0.35$ reference budget, Full ACE achieves 94.30% POPE-Adversarial accuracy, compared with 92.27% and 92.25% for uniform and shuffled allocation. Removing S, V, or $(1 - \xi)$ produces 93.30%, 94.07%, and 94.13% accuracy, respectively. These controlled comparisons identify reliability-based allocation as an important source of ACE's performance advantage over unstructured amplification. Component-level sensitivity and paired uncertainty are analyzed in Appendix F.

6 Limitations

ACE requires architecture-specific visual-unit mappings, access to per-source attention contributions, and additional validation and decoder-replay computation. The reference implementation uses eager attention kernels. Counter-context acceptance measures feature preservation outside the edited support, while fixed CCIB coverage and concept exclusion may affect different query categories. Reliability scores are label-free selection proxies, whereas AFE measures single-unit deletion effects. Our mechanistic analysis focuses on first-answer decisions with explicit competing hypotheses; extending these diagnostics to open-ended and temporal reasoning requires additional protocols.

7 Conclusion

We characterize visual hallucination through layer-wise evidence competition, identifying strict silencing as a failure in which robust intermediate factual dominance is overturned before terminal prediction. To address this failure, ACE introduces controlled counter-context validation and anchor-based sustained preservation, selectively amplifying current visual contributions along the original-image decoder suffix. Across four MLLM backbones, ACE improves POPE-Adversarial accuracy by 6.37–8.97 percentage points and reduces CHAIR_s by 3.9–5.0 points. Functional diagnostics and controlled preservation experiments further demonstrate the value of reliability-based evidence identification and sustained selective intervention. Together, these results establish sustained visual-evidence influence as a practical inference-time target for mitigating multimodal hallucination.

ETHICS STATEMENT

This work investigates hallucination in multimodal large language models and proposes an inference-time method to improve factual reliability. Experiments use established benchmarks and controlled counter-context interventions without modifying benchmark annotations or training the evaluated models. ACE does not guarantee factual correctness, and its reliability estimates should not be interpreted as safety guarantees. Upon acceptance, we will release the implementation and associated reproducibility resources, with counter-context assets and benchmark materials distributed in accordance with their applicable licenses and privacy requirements.

AI USE STATEMENT

Generative AI tools, primarily OpenAI ChatGPT, were used to assist with research development and manuscript preparation. These tools supported discussions of the problem formulation, theoretical framework, mathematical derivations, methodological design, and experimental protocols. They also assisted with literature searches and verification, code development, metric checking, English editing, and LaTeX preparation. AI assistance in figure preparation was limited to visualization design, plotting, and presentation of experimental findings. The authors reviewed and verified AI-assisted content, mathematical arguments, code, citations, and analyses, made all final scientific decisions, and take full responsibility for the accuracy, originality, and integrity of this work.

REPRODUCIBILITY STATEMENT

We provide detailed methodological and experimental information to facilitate the reproduction of our results. The appendix documents the diagnostic formulation and mathematical derivations, counter-context construction, visual-unit mappings, reliability estimation, inference algorithm, and decoder-suffix replay procedure. It further specifies benchmark protocols, fixed experimental configurations, statistical aggregation, paired evaluations, mechanistic ablations, sensitivity analyses, matched-budget controls, and inference-efficiency measurements. Upon acceptance, we will publicly release the ACE source code, backbone-specific implementations, the curated Counter-Context Intervention Bank (CCIB), subject to redistribution permissions, and evaluation artifacts, including experimental configurations, sample identifiers, derived evaluation data, and scripts required to reproduce the reported results, tables, and figures.

REFERENCES

Samira Abnar and Willem Zuidema. Quantifying attention flow in transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4190–4197. Association for Computational Linguistics, 2020. doi: 10.18653/v1/2020.acl-main.385. URL https://aclanthology.org/2020.acl-main.385/.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1723.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, Mei Li, Kaixin Li, Zicheng Lin, Junyang Lin, Xuejing Liu, Jiawei Liu, Chenglong Liu, Yang Liu, Dayiheng Liu, Shixuan Liu, Dunjie Lu, Ruilin Luo, Chenxu Lv, Rui Men, Lingchen Meng, Xuancheng Ren, Xingzhang Ren, Sibo Song, Yuchong Sun, Jun Tang, Jianhong Tu, Jianqiang Wan, Peng Wang, Pengfei Wang, Qiuyue Wang, Yuxuan Wang, Tianbao Xie, Yiheng Xu, Haiyang Xu, Jin Xu, Zhibo Yang, Mingkun Yang, Jianxin Yang, An Yang, Bowen Yu, Fei Zhang, Hang Zhang, Xi Zhang, Bo Zheng, Humen Zhong, Jingren Zhou, Fan Zhou, Jing Zhou, Yuanzhi Zhu, and Ke Zhu. Qwen3-VL technical report. *arXiv preprint arXiv:2511.21631*, 2025a. URL https://arxiv.org/abs/2511.21631.

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*, 2025b. URL https://arxiv.org/abs/2502.13923.

Nora Belrose, Zach Furman, Logan Smith, Danny Halawi, Igor Ostrovsky, Lev McKinney, Stella Biderman, and Jacob Steinhardt. Eliciting latent predictions from transformers with the tuned lens. *arXiv preprint arXiv:2303.08112*, 2023. URL https://arxiv.org/abs/2303.08112.

Jiale Chang, Ying Li, Siliang Tang, and Yueting Zhuang. AHEAD: Attention head energy-aware dynamics for hallucination mitigation in MLLMs. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 8728–8739. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.425. URL https://aclanthology.org/2026.findings-acl.425/.

Hila Chefer, Shir Gur, and Lior Wolf. Transformer interpretability beyond attention visualization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 782–791, 2021. doi: 10.1109/CVPR46437.2021.00084.

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.

Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James R. Glass, and Pengcheng He. DoLa: Decoding by contrasting layers improves factuality in large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/edc36117f795ca52a0cbf6a7b3882859-Abstract-Conference.html.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven 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.

Tiantian Dang, Chao Bi, Shufan Shen, Jinzhe Liu, Qingming Huang, and Shuhui Wang. Locate-then-sparsify: Attribution guided sparse strategy for visual hallucination mitigation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18251–18260, 2026.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/0b408293619f725fd30162af057e531a-Abstract-Conference.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. doi: 10.1109/CVPR52733.2024.01363.

Yichen Guo, Kai Tang, Jinhao You, Fenglai Lin, Yiding Sun, Dongxu Zhang, Wenya Wang, Lin William Cong, and Shanghang Zhang. FADE: Mitigating hallucinations by reducing language-prior dominance in large vision-language models. *arXiv preprint arXiv:2606.29431*, 2026. doi: 10.48550/arXiv.2606.29431. URL https://arxiv.org/abs/2606.29431.

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, 2024a. doi: 10.1109/CVPR52733.2024.01274.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. Visual hallucinations of multi-modal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 9614–9631. Association for Computational Linguistics, 2024b. doi: 10.18653/v1/2024.findings-acl.573. URL https://aclanthology.org/2024.findings-acl.573/.

Sarthak Jain and Byron C. Wallace. Attention is not explanation. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 3543–3556. Association for Computational Linguistics, 2019. doi: 10.18653/v1/N19-1357. URL https://aclanthology.org/N19-1357/.

Samyak Jha and Junho Kim. CAPA: Contribution-aware pruning and FFN approximation for efficient large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2026*, pp. 27543–27558. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.findings-acl.1371. URL https://aclanthology.org/2026.findings-acl.1371/.

Junyang Ji, Qifan Liu, Wenming Yang, and Zhihai He. CausalLens: Sensitivity-guided multi-head causal intervention for hallucination mitigation in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 4199–4209, 2026.

Nicholas Jiang, Amil Dravid, Alexei A. Efros, and Yossi Gandelsman. Vision transformers don't need trained registers. In *Advances in Neural Information Processing Systems*, 2025. doi: 10.52202/085713-1891.

Zhehan Kan, Xin Li, Yanlin Liu, Xiaochen Yang, Xinghua Jiang, Yinsong Liu, Deqiang Jiang, Xing Sun, Qingmin Liao, and Wenming Yang. RAR: Reversing Visual Attention Re-Sinking for Unlocking Potential in Multimodal Large Language Models. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/42a1f9ef76af656f3fe33580ab794351-Abstract-Conference.html.

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. doi: 10.1109/CVPR52733.2024.01316.

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

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. SEED-Bench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13299–13308, 2024. doi: 10.1109/CVPR52733.2024.01263.

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.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model. In *Advances in Neural Information Processing Systems*, volume 36, 2023b. doi: 10.52202/075280-1797.

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

Yuxuan Liang, Fan Shi, Rui Zhu, Xu Li, Xiaolei Chen, Zhe Liu, Bin Li, and Xiangyang Xue. Envision, attend, then respond: Counterfactual hallucination mitigation in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18261–18272, 2026.

Yangguang Lin, Quan Fang, Yufei Li, Jiachen Sun, Junyu Gao, and Jitao Sang. HulluEdit: Single-pass evidence-consistent subspace editing for mitigating hallucinations in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11086–11095, 2026.

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 *International Conference on Learning Representations*, 2024a. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/fed9263d7f1086e735904ff690527a53-Abstract-Conference.html.

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.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to images: A training-free method for alleviating hallucination in LVLMs. In *European Conference on Computer Vision*. Springer, 2024b. doi: 10.1007/978-3-031-73010-8_8.

Shuliang Liu, Songbo Yang, Dong Fang, Sihang Jia, Yuqi Tang, Lingfeng Su, Ruoshui Peng, Yibo Yan, Xin Zou, and Xuming Hu. Vision-language introspection: Mitigating overconfident hallucinations in MLLMs via interpretable bi-causal steering. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 38518–38543. Association for Computational Linguistics, 2026. doi: 10.18653/v1/2026.acl-long.1784. URL https://aclanthology.org/2026.acl-long.1784/.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. MMBench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*, pp. 216–233. Springer, 2024c. doi: 10.1007/978-3-031-72658-3_13.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In *Advances in Neural Information Processing Systems*, volume 35, 2022. doi: 10.52202/068431-1262.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mahmoud Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. DINOv2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. URL https://openreview.net/forum?id=a68SUt6zFt.

Nina Rimsky, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15504–15522. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.828. URL https://aclanthology.org/2024.acl-long.828/.

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/.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. Eyes wide shut? exploring the visual shortcomings of multimodal LLMs. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9568–9578, 2024. doi: 10.1109/CVPR52733.2024.00914.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. MLLM Can See? Dynamic Correction Decoding for Hallucination Mitigation. In *International Conference on Learning Representations*, 2025a. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/24079b91da7257cb78805262996152b8-Abstract-Conference.html.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. AMBER: An LLM-free multi-dimensional benchmark for MLLMs hallucination evaluation. *arXiv preprint arXiv:2311.07397*, 2023. URL https://arxiv.org/abs/2311.07397.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b. URL https://arxiv.org/abs/2508.18265.

Zichuan Wang, Songlin Yang, Bo Peng, Zhenchen Tang, Yang Li, Beibei Dong, and Jing Dong. Same attention, different truths: Put logit-lens over visual attention to detect and mitigate LVLM object hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25315–25325, 2026a.

Zihu Wang, Boxun Xu, Yuxuan Xia, and Peng Li. VEGAS: Mitigating hallucinations in large vision-language models via vision-encoder attention guided adaptive steering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Findings*, pp. 9055–9064, 2026b.

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*, 67:220105, 2024. doi: 10.1007/s11432-024-4251-x.

Zhihan Yin, Jianxin Liang, Yueqian Wang, Yifeng Yao, Huishuai Zhang, and Dongyan Zhao. FREAK: A fine-grained hallucination evaluation benchmark for advanced MLLMs. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/51c06741bacd0987a9f1ed9dc243b4db-Abstract-Conference.html.

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, 2024a.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. MM-vet: Evaluating large multimodal models for integrated capabilities. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 57730–57754. PMLR, 2024b. URL https://proceedings.mlr.press/v235/yu24o.html.

Zihao Yue, Liang Zhang, and Qin Jin. Less is more: Mitigating multimodal hallucination from an EOS decision perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11766–11781. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.633. URL https://aclanthology.org/2024.acl-long.633/.

Chengsheng Zhang, Chenghao Sun, Xinyan Jiang, Wei Li, and Xinmei Tian. Prefill-time intervention for mitigating hallucination in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25293–25303, 2026.

Shaolei Zhang, Tian Yu, and Yang Feng. TruthX: Alleviating hallucinations by editing large language models in truthful space. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 8908–8949. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.acl-long.483. URL https://aclanthology.org/2024.acl-long.483/.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/50623630a2372839c078474efa6c0cb8-Abstract-Conference.html.

Supplementary Material

A Diagnostic Formulation and Robustness

This section provides the diagnostic details underlying Sec. 3. These quantities are used only to characterize layer-wise evidence competition and are never used as supervision by ACE during inference.

A.1 Lexical Hypothesis Scoring

The main text defines the factual competition margin at generation position t as

$$m_{\ell,t} = s_{\ell,t}(y_f) - s_{\ell,t}(y_c).$$

For POPE, the mechanistic analysis concerns the first-answer decision at the final prompt position t_0.

For each hypothesis y, we construct a model-specific lexical set $\mathcal{T}(y)$ containing tokenizer-level first-token realizations accepted by the benchmark answer convention. This accounts for tokenizer-dependent variations such as leading whitespace and capitalization while leaving the benchmark parser unchanged. The lexical sets are fixed for each tokenizer before evaluation and are not selected on a per-example basis.

Given the normalized intermediate readout

$$\mathbf{z}_{\ell,t} = W_U \,\mathrm{Norm}(h_{\ell,t}),$$

the hypothesis score is

$$s_{\ell,t}(y) = \log \sum_{u \in \mathcal{T}(y)} \exp z_{\ell,t}(u). \tag{11}$$

For a single lexical realization, $\mathcal{T}(y)$ reduces to a singleton and Eq. 11 becomes the corresponding output logit.

The default robust-dominance threshold in the main experiments is

$$\delta = \log \frac{0.6}{0.4} \approx 0.405465.$$

To interpret this value, define the probability obtained after normalizing only between the two diagnostic hypotheses:

$$p^{\text{bin}}_{f,\ell} = \frac{\exp s_{\ell,t_0}(y_f)}{\exp s_{\ell,t_0}(y_f) + \exp s_{\ell,t_0}(y_c)} = \sigma(m_{\ell,t_0}). \tag{12}$$

Thus, $m_{\ell,t_0} \geq \delta$ if and only if $p^{\text{bin}}_{f,\ell} \geq 0.6$. This is a pairwise diagnostic probability and should not be interpreted as the probability of the factual hypothesis over the full output vocabulary.

A.2 Robust Dominance, Reversal, and Strict Silencing

The robust competition state associated with the margin in Eq. 1 is

$$r_{\ell,t} = \begin{cases} +1, & m_{\ell,t} \geq \delta, \\ -1, & m_{\ell,t} \leq -\delta, \\ 0, & |m_{\ell,t}| < \delta. \end{cases} \tag{13}$$

States with $r_{\ell,t} = 0$ form the weak band and are removed before counting reversals. Let $(r_{\ell_1,t}, \ldots, r_{\ell_M,t})$ denote the remaining ordered nonzero states. A robust dominance reversal occurs whenever

$$r_{\ell_k,t} \neq r_{\ell_{k+1},t}.$$

The reversal count is therefore

$$N_{\text{rev}} = \sum_{k=1}^{M-1} \mathbb{I}\left[r_{\ell_k,t} \neq r_{\ell_{k+1},t}\right]. \tag{14}$$

This symmetric definition counts both factual-to-competing and competing-to-factual transitions while suppressing small oscillations around zero.

The trajectory-level silencing event in Eq. 2 becomes an empirical failure set only after intersecting it with errors produced by the unmodified model:

$$\mathcal{E}^{\text{raw}}_{\text{sil}} = \left\{ i \in \mathcal{E}^{\text{raw}}_{\text{err}} : \max_{\ell<L} m^i_{\ell,t_0} \geq \delta,\ m^i_{L,t_0} \leq -\delta \right\}. \tag{15}$$

The strongest pre-terminal factual state is

$$\ell^\star = \arg\max_{\ell<L} m_{\ell,t_0}. \tag{16}$$

Ties are resolved in favor of the shallower layer. Importantly, $\ell^\star$ is used only for diagnostic analysis and is not available to ACE during inference.

A.3 Terminal-Takeover Derivation

The terminal-takeover identity in Eq. 3 follows directly from the layer-wise margin increments

$$\Delta m_{\ell,t} = m_{\ell+1,t} - m_{\ell,t}.$$

Starting from any layer $\ell_0 < L$,

$$\begin{aligned} m_{L,t} &= m_{\ell_0,t} + \sum_{k=\ell_0}^{L-1} (m_{k+1,t} - m_{k,t}) \\ &= m_{\ell_0,t} + \sum_{k=\ell_0}^{L-1} \Delta m_{k,t}. \end{aligned} \tag{17}$$

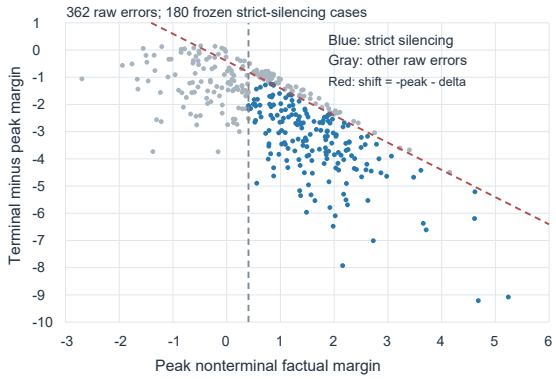


Figure 5: **Diagnostic terminal-takeover geometry.** Peak factual margin versus downstream shift for all 362 Qwen3 raw errors. Points use the same frozen cohort as the main analysis. The red boundary $y = -(x + \delta)$ is analytically determined by the strict-silencing criterion.

Setting $\ell_0 = \ell^\star$ gives

$$m_{L,t} = m_{\ell^\star,t} + \sum_{k=\ell^\star}^{L-1} \Delta m_{k,t}. \tag{18}$$

Therefore,

$$m_{L,t} \le -\delta \iff \sum_{k=\ell^\star}^{L-1} \Delta m_{k,t} \le -\left(m_{\ell^\star,t} + \delta\right), \tag{19}$$

which is the boundary visualized in Fig. 5.

Equation 18 is an exact telescoping identity for the aggregated lexical hypothesis score, providing the analytical boundary used to characterize terminal takeover.

For interpretation, consider the special case in which both hypotheses are represented by single output tokens. Let

$$\widetilde{h}_{\ell,t} = \mathrm{Norm}(h_{\ell,t}),$$

and let w_f and w_c denote the corresponding rows of the output head. With $\Delta w = w_f - w_c$,

$$m_{\ell,t} = \Delta w^\top \widetilde{h}_{\ell,t}, \qquad \Delta m_{\ell,t} = \Delta w^\top \left(\widetilde{h}_{\ell+1,t} - \widetilde{h}_{\ell,t}\right). \tag{20}$$

Eq. 20 provides a fixed-output- direction interpretation of the margin change but is not an additional assumption required by ACE.

A.4 Robustness to the Dominance Threshold

For $p \in \{0.55, 0.60, 0.65, 0.70\}$, define $\delta_p = \log(p/(1-p))$. Predictions and trajectories remain fixed. For each threshold, a raw-error case enters the strict subset iff its nonterminal peak is at least δ_p and its terminal margin is at most $-\delta_p$. Rescue additionally requires the stored ACE prediction to be correct and its terminal margin to reach δ_p. Fallback cases are retained. Reversals are sign changes after removing states with $|m| < \delta_p$. Table 4 reports integer counts; it is not an average over repeated experiments.

The strict sets and their rescue counts are nested as the threshold rises. The default row reuses 180 raw strict cases and 52 strict rescues. The rescue fraction decreases for these fixed trajectories.

B ACE Implementation Details

This section specifies the operations and replay boundary underlying Sec. 4. The organization follows the same *Challenge*, *Validate*, *Locate*, and *Preserve* sequence used in the main text.

Table 4: **Post hoc threshold sensitivity on the fixed Qwen3 cohort.** Raw and ACE trajectories, predictions, gates and anchors are fixed across rows. Correct- and error-reversal denominators are 2,638 and 362; strict-rescue denominators are the threshold-specific strict counts. These are integer counts, not repeated-run averages.

p	δ_p	Strict/errors	Correct rev.	Error rev.	Strict rescue
0.55	0.201	243/362	42.6%	73.5%	110/243 (45.3%)
0.60	0.405	180/362	30.9%	66.3%	52/180 (28.9%)
0.65	0.619	150/362	23.7%	56.6%	27/150 (18.0%)
0.70	0.847	122/362	16.3%	48.9%	7/122 (5.7%)

B.1 Visual-Conditioned Units and Projected Contributions

A unit $j \in \mathcal{I}_V$ is a native decoder input position carrying an image-derived embedding. Original, counter-context, and null views share image dimensions, crop geometry, packing order, and position IDs. The visual-unit mask excludes padding, image delimiters, and learned newline tokens; valid nonvisual context positions remain in $\mathcal{I}_{\mathrm{ctx}}$ for the decision-coupling denominator.

Attention interface. For each query head, ACE uses its attention coefficient and associated value vector, expanding the key/value-head assignment logically in grouped-query attention. Multiplication by the corresponding output projection gives $c_{\ell,t,j}$ in Eq. 4. The reference hook applies $1 + \alpha\omega_{\ell,j}$ to visual-source attention-weighted values before the output projection, without renormalizing attention weights. Linearity of the projection makes this equivalent to Eq. 10. Output-projection bias is added once and is not attributed to any source. The hook acts at the current answer-decision position; it does not rescale the whole KV cache or modify earlier prompt queries. Native residual and MLP operations remain unchanged. Source contributions can be accumulated in chunks; a tensor over every query, source, and hidden dimension is unnecessary.

Geometric support. The common image domain is a 16×16 grid over the entire original image. Let P_p be grid cell p and R_u the inverse-mapped footprint of native patch u. Resize, crop, tile, padding, and packing transforms are retained explicitly. For patches with nonzero valid-image area, define the overlap weights

$$G_{u,p} = \frac{|R_u \cap P_p|}{|R_u \cap I|}, \qquad A_{j,p} = \frac{\sum_u b_{j,u} G_{u,p}}{\sum_u b_{j,u}},$$

where $b_{j,u} \geq 0$ records uniform grouping or spatial interpolation weights. Fully padded patches contribute zero weight. Thus,

$$\sum_p A_{j,p} = 1. \tag{21}$$

These weights specify geometric provenance. They are not learned projector weights, a causal attribution map, or the full receptive field after globally mixing vision attention. Overlapping views retain separate decoder units and their own normalized supports.

Qwen3-VL. The visual frontend uses 16×16 patches and a spatial merge factor of two. Each merged unit inherits the union of its four native patch footprints, mapped through the recorded resize transform. The packed visual indices are shared by the original and edited views. DeepStack visual additions remain native and unchanged; ACE intervenes on the source-specific attention contributions defined above, not on every possible visual pathway.[1]

Qwen2.5-VL. The native patch size is 14 and the spatial merge factor is two. The support of a merged unit is the uniform union of its 2×2 source patches. Any vision-window or merger ordering is inverted before associating the final packed unit with its image coordinates. Consequently, the

[1] Configuration and reference implementation: `https://huggingface.co/Qwen/Qwen3-VL-8B-Instruct/blob/main/config.json`; `https://github.com/huggingface/transformers/blob/v4.57.1/src/transformers/models/qwen3_vl/modeling_qwen3_vl.py`.

mapping follows source indices rather than assuming that an intermediate window order is the final raster order.[2]

InternVL3.5. Each 448×448 tile produces a 32×32 grid of 14-pixel patches. After removal of the class token, pixel shuffle with ratio 0.5 groups 2×2 patches, producing 256 projected units per tile. Each unit is mapped through that tile's crop transform. When a global thumbnail is present, its 256 units form a separate global view; their supports are not assigned to a single local tile. The pixel-shuffle permutation and the concatenated tile order are retained explicitly. [3]

LLaVA-OneVision. The selected SigLIP frontend uses a 384-pixel image input and 14-pixel patches. The mapping follows the base view and any-resolution tiles through spatial packing, unpadding, and any subsequent bilinear resize. For an interpolated unit, $b_{j,u}$ is the nonnegative interpolation weight on its source grid cells. Learned newline positions are excluded from $\mathcal{I}_V$. The number of packed visual positions depends on the native preprocessing geometry and is not assumed to equal 576. [4]

B.2 Counter-Context Intervention Bank

The Counter-Context Intervention Bank (CCIB) is constructed offline before evaluation and remains fixed throughout all experiments. The bank configuration contains $K = 500$ candidate visual assets. For asset k, we store

$$\mathcal{B}_k = (O_k, M_k, e_k, \mathcal{A}_k)\,, \tag{22}$$

where O_k is the visual asset, M_k its foreground mask, e_k its normalized visual-semantic embedding, and $\mathcal{A}_k$ a set of semantic metadata and aliases.

The bank is used only to construct a contextual challenge. No benchmark answer label, target output, or model prediction is used to construct or update the bank during evaluation. Before retrieval for query q, assets matching an explicit concept in the task text are excluded using the fixed CCIB metadata and aliases. Appendix B.3 specifies the matching rule and the empty-target branch. This rule excludes matched queried concepts; it is not a general detector of all objects present in the image.

The embedding e_k is cached once during offline construction. At inference time, only retrieval, placement, and validity checking are performed. The same fixed bank is shared across all test examples.

Validity encoder and feature interface. Candidate checking uses a separate frozen DINOv2-S/14 encoder Φ (Oquab et al., 2024), shared across backbones. Its input is RGB resized directly to 224×224 without a center crop, scaled to $[0, 1]$, and normalized by mean $(0.485, 0.456, 0.406)$ and standard deviation $(0.229, 0.224, 0.225)$. The direct resize intentionally replaces the checkpoint processor's default center crop, retaining the full image domain. Final normalized patch outputs, excluding the class token, give a 16×16 grid; each patch vector is L2-normalized to form g_p.[5] The bank vector e_k is the normalized foreground-mask-weighted mean of these patch features for asset k. Both v_C and v_C^{cf} are normalized means over the same cell indices $C_\gamma(I, q)$, selected once from the original relevance map. The lowest-relevance fraction contains $\lceil \gamma\, 256 \rceil$ cells, with raster order breaking ties. Features of a zero-norm pooled vector are treated as invalid for candidate acceptance.

Search and pass boundary. Original features $\Phi(I)$ are computed once and cached per request; bank embeddings are cached offline. After semantic exclusion and the fixed d_k interval, candidates

[2] https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct/blob/main/config.json.

[3] Configuration and visual extraction: https://huggingface.co/OpenGVLab/InternVL3_5-8B/blob/main/config.json; https://huggingface.co/OpenGVLab/InternVL3_5-8B/blob/main/modeling_internvl_chat.py.

[4] Configuration and spatial packing: https://huggingface.co/lmms-lab/llava-onevision-qwen2-7b-ov/blob/main/config.json; https://github.com/LLaVA-VL/LLaVA-NeXT/blob/main/llava/model/llava_arch.py.

[5] Validity-encoder configuration and normalization: https://huggingface.co/facebook/dinov2-small/blob/main/config.json; https://huggingface.co/facebook/dinov2-small/blob/main/preprocessor_config.json.

are ordered by distance to the midpoint of that interval, with bank ID breaking ties. At most eight candidates undergo placement and validity checking. Candidate scale is fixed by bank metadata; placement uses the allowed image-grid positions and breaks ties in raster order. A grid cell intersecting an edited mask belongs to the edit support for both preservation and exposure. Every shortlisted candidate is checked before choosing the admissible one with maximum D_C (ties use bank ID). "No admissible candidate" means none in this bounded search, not an exhaustive search over all bank assets and placements.

Each edited candidate requires only image composition and a forward through Φ to calculate P_{pres} and D_C; it does not require a language-decoder prefill. Only the selected probe is then processed by the backbone's native visual frontend and language decoder to obtain layer-wise values for $S_{\ell,j}$. Candidate features g_p are never substituted for native MLLM visual embeddings. Selection therefore precedes the reliability pass without assuming that the selected probe's decoder states already exist.

B.3 Query-Conditioned Relevance and Candidate Placement

ACE constructs the counter-context from quantities obtained during the original-image prefill at t_0.

Task-specific query interface. The runtime query q is the task input supplied to the evaluated MLLM, serialized with the backbone's native image and chat template. The task-text component contains the POPE object-existence question, the FREAK question together with its presented answer-option texts when applicable, or the original MM-Vet question without an additional task-specific instruction or answer prefix. For CHAIR, the task text is exactly `Describe the image in detail.` The original-image prefill conditions on this complete input before the first answer token is selected. Thus, an open-ended captioning request still defines a query for the relevance map; it does not require a preselected object label. No preliminary caption is generated to construct the probe.

Explicit-concept exclusion. The exclusion interface uses only the task text and the fixed CCIB alias dictionary. Text and aliases are case-folded; non-alphanumeric characters are replaced by spaces and consecutive spaces are collapsed. An alias matches only as a complete contiguous token sequence. Inflected forms and synonyms match only when explicitly stored in the dictionary. Every concept with a matching alias is included in the exclusion set, and an asset is removed when its metadata contain any such concept. For multiple-choice inputs, matching covers the question and all presented option texts equally, without selecting an option. Native chat delimiters and system-template text are not used for alias matching. This is a lexical alias rule, not sentence-embedding matching.

Queries without an explicit target. When the exclusion set is empty, including for the generic CHAIR prompt above, no asset is removed by target-concept exclusion. No target is inferred from a generated answer, reference caption, ground-truth annotation, or auxiliary object detector. The relevance map, bounded candidate distance, placement rule, preservation test, contextual-displacement test, and fallback remain unchanged. Consequently, this branch does not guarantee that an asset is absent from the original scene. It instead subjects that asset to the same query-conditioned placement and probe-validity constraints.

We first aggregate each unit's normalized contribution magnitude across non-terminal decoder layers:

$$\overline{R}_j = \frac{1}{L-1} \sum_{\ell=1}^{L-1} \frac{\|c_{\ell,t_0,j}\|_2}{\sum_{i \in \mathcal{I}_{\text{ctx}}} \|c_{\ell,t_0,i}\|_2 + \epsilon}. \tag{23}$$

Using the spatial support in Eq. 21, the normalized image-level relevance map is

$$\pi_p(I,q) = \frac{\sum_{j \in \mathcal{I}_V} \overline{R}_j A_{j,p}}{\sum_{p'} \sum_{j \in \mathcal{I}_V} \overline{R}_j A_{j,p'} + \epsilon}. \tag{24}$$

This map is a contribution-magnitude-based estimate of spatial relevance, not a ground-truth annotation of query-relevant content. It is used only for counter-context construction, not as an additional evidence score during answer generation.

Let $C_\gamma(I, q)$ denote the lowest-relevance fraction of the image under $\pi_p(I, q)$, and let $v_C(I, q)$ denote the corresponding mean normalized visual feature. For CCIB candidate k, semantic displacement from the local context is

$$d_k = 1 - \cos\left(v_C(I, q), e_k\right). \tag{25}$$

ACE uses a bounded semantic range rather than selecting the most distant candidate. This removes candidates that are too close to the existing context while avoiding extreme contextual perturbations.

For candidate k, let $\Omega_k(r)$ denote its spatial footprint when placed at location r. Its query-aware placement is

$$r_k^\star = \arg\min_r \sum_{p \in \Omega_k(r)} \pi_p(I, q), \tag{26}$$

subject to the complete candidate mask remaining inside the valid image region. Minimizing this estimated relevance mass encourages low-risk placement but imposes no absolute upper bound on that mass. If the map underestimates a relevant region, placement can still obscure useful content in the probe.

B.4 Probe Validity and Edit Exposure

Let $\Omega^\star$ denote the support of the final counter-context edit, and let $g_p(\cdot)$ denote the local visual feature at image location p. Relevance-weighted feature preservation outside the edited support is measured by

$$P_{\text{pres}} = \frac{\sum_{p \notin \Omega^\star} \pi_p(I, q) \cos\left(g_p(I), g_p(I^{\text{cf}})\right)}{\sum_{p \notin \Omega^\star} \pi_p(I, q) + \epsilon}. \tag{27}$$

P_{pres} is a regularized weighted average of feature similarities over $p \notin \Omega^\star$, with weights estimated from the original input. It does not inspect content within $\Omega^\star$, and a high average does not impose a pointwise similarity bound even outside the edit. Thus, the score cannot certify preservation of every query-relevant cue. Low-relevance placement reduces estimated exposure to such cues but depends on the quality of the relevance map.

The contextual displacement induced by the edit is

$$D_C = 1 - \cos\left(v_C, v_C^{\text{cf}}\right), \tag{28}$$

where the two context vectors are computed over the corresponding low-relevance support.

A candidate is accepted only when

$$P_{\text{pres}} \geq \tau_{\text{pres}}, \qquad D_C^{\min} \leq D_C \leq D_C^{\max}. \tag{29}$$

Here, probe validity means admissibility under Eq. 29; it is not a certificate of complete semantic or factual preservation. Among admissible candidates, ACE chooses the one with the largest contextual displacement. If none satisfies these criteria, ACE falls back to standard decoding.

Although placement targets a low-relevance image region, one decoder-facing visual unit may aggregate information from multiple locations. We therefore define its direct edit exposure as

$$\xi_j = \sum_{p \in \Omega^\star} A_{j,p}. \tag{30}$$

Since the support map is normalized, $\xi_j \in [0, 1]$. The factor $(1 - \xi_j)$ in Eq. 8 therefore provides a continuous discount rather than a hard removal of units influenced by the intervention. This discount does not certify that the probe retains every relevant cue within the edited support.

B.5 Null-Visual References

The feature $x_j^V(I, q)$ in Eq. 7 is the embedding supplied at visual input position j after the native vision encoder, merger/projector, and spatial packing, before the language decoder. The query index identifies the paired request; this feature does not contain a query-conditioned language-decoder state.

It is distinct from both the validity feature g_p and the layer-wise value $v_{\ell,j}$ used for counter-context stability.

The null set contains three images at the original resolution: a constant image using the original per-channel mean, a Gaussian-blurred image with $\sigma = 0.1 \min(H, W)$ and a 3σ kernel radius, and independent Gaussian noise matched to the original per-channel mean and standard deviation and clipped to $[0, 1]$. Reflection padding is used for blur. Noise is fixed per image using the low 32 bits of its RGB pixel SHA-256 digest as the PCG64 seed. The same null images are reused across paired configurations. None depends on an answer label.

Each null image uses the original crop/resize geometry and native visual frontend, preserving the unit correspondence. The three visual passes are sequential at batch size one; they require no language-model prefill under this feature definition. Original-image embeddings are reused from the raw pass. ACE takes the maximum null similarity for each unit and computes V_j in FP32. All three online visual passes contribute to TTFT. If no candidate is admissible, they are skipped; if evidence fallback occurs after validation, their cost is retained.

B.6 Fixed Intervention Configuration

Unless otherwise specified, the following reference configuration is shared across the evaluated backbones and benchmarks:

$$\gamma = 0.20, \qquad \tau_{\text{pres}} = 0.90, \qquad d_k \in [0.30, 0.65], \qquad D_C \in [0.25, 0.60], \qquad \alpha = 0.35. \quad (31)$$

Default-setting policy. The constants in Eq. 31 define fixed design choices rather than optima from a validation-set search. No dataset-driven hyperparameter search is part of this reference protocol, and the defaults are not adjusted for individual test examples. The choice $\gamma = 0.20$ restricts the context descriptor to the lowest-relevance fifth of the image grid. The threshold $\tau_{\text{pres}} = 0.90$ imposes a high outside-edit feature-similarity requirement. The interval $d_k \in [0.30, 0.65]$ constrains asset retrieval before composition, whereas $D_C \in [0.25, 0.60]$ constrains the resulting contextual change after composition; the two intervals serve different operations and are not interchangeable. These are operational design rationales, not claims of empirically optimal threshold values.

The main comparison uses $\alpha = 0.35$ as a common reference strength. Since $\omega_{\ell,j} \in [0, 1]$, this choice bounds the per-source Preserve multiplier $1 + \alpha\omega_{\ell,j}$ between 1 and 1.35. This is a bound on each multiplier, not on the net residual perturbation or on factual improvement. The preservation-strength sweep in Appendix D.5 evaluates sensitivity separately; the reference setting is not presented as the optimum of that sweep.

Separate amplitude matching. The calibration in Appendix F serves a different purpose: it determines a variant-specific amplitude coefficient for matching a given Full ACE reference budget. It does not select γ, τ_{pres}, either distance interval, or the default reference strength. The main comparison and the matched-budget comparison therefore retain distinct parameter-selection roles.

B.7 ACE Inference Algorithm

Algorithm 1 summarizes the complete inference procedure. The ordering is important: the counter-context probe is completed and discarded before the first answer token is selected.

The selected counter-context branch is evaluated once for reliability estimation and is not maintained as an autoregressive generation stream. Candidate validity checking and the three null visual passes are separate online operations. They are charged explicitly to TTFT, even though neither requires a candidate-level or null-level language-decoder prefill.

B.8 Replay Boundary and KV-Cache Reuse

We treat decoder layer ℓ as the block whose attention contribution $c_{\ell,t,j}$ is written into the representation produced by that block. The anchor $\ell^\dagger$ is therefore the first block at which ACE preservation is enabled.

Algorithm 1 Adaptive Counter-context Evidence (ACE)

Require: Image I, query q, MLLM f, fixed CCIB $\mathcal{B}$, preservation strength α
Ensure: Generated response $\hat{\mathbf{y}}$

Run the original-image prefill on (I, q) through t_0; retain the original states and caches without selecting an answer token
Collect $c_{\ell,t_0,j}$ and construct $\pi(I, q)$
Cache $\Phi(I)$; search at most eight candidates using visual-only checks in Eqs. 25–29
if no candidate is admissible **then**
    return standard decoding on (I, q)
end if
Select the admissible candidate with maximum allowed D_C and run its native visual frontend and language-decoder prefill
Run the three null visual frontends (no language-decoder prefill); compute V_j and ξ_j for every visual unit
for each layer $\ell = 1, \ldots, L$ and visual unit j **do**
    Compute $R_{\ell,t_0,j}$ and $S_{\ell,j}$
    $\omega_{\ell,j} \leftarrow (1 - \xi_j) S_{\ell,j} V_j$; $Q_{\ell,t_0,j} \leftarrow R_{\ell,t_0,j} \omega_{\ell,j}$
end for
$C^E_\ell \leftarrow \sum_{j \in \mathcal{I}_V} Q_{\ell,t_0,j}$ for $\ell < L$; $\ell^\dagger \leftarrow \arg\max_{\ell<L} C^E_\ell$
if $\max_{\ell<L} C^E_\ell$ is numerically negligible **then**
    return standard decoding on (I, q)
end if
Discard counter-context states, outputs, and logits; retain the estimated original-unit reliability gates
Restore the original-image input to block $\ell^\dagger$ at t_0
for $\ell = \ell^\dagger, \ldots, L$ **do**
    Replay block ℓ at t_0, applying $\widetilde{c}_{\ell,t_0,j} = (1 + \alpha\omega_{\ell,j}) c_{\ell,t_0,j}$ to visual contributions inside the block
    Recompute states and cache entries that depend on intervened representations; reuse unaffected entries
end for
Select the first answer token from the replayed original-image state
while generation is not finished **do**
    Decode the next token on the original-image trajectory, applying Eq. 10 within each block $\ell \geq \ell^\dagger$ before the token is selected
end while
return $\hat{\mathbf{y}}$

Before the first answer token is selected, ACE returns to the original-image computation and replays the decoder suffix beginning at $\ell^\dagger$. All downstream quantities that depend on an altered representation are recomputed.

Under causal masking, prompt positions preceding t_0 cannot depend on the representation at the final prompt position. Their unchanged cached quantities can therefore be reused. At t_0, ACE recomputes downstream hidden states and cache entries whose key or value projections depend on an intervened representation, while cache entries determined entirely by unchanged prefix representations are retained.

For later answer tokens, the reliability gates $\omega_{\ell,j}$ remain fixed from the validation stage, whereas the contributions $c_{\ell,t,j}$ are recomputed from the evolving original-image decoding state. This distinction makes the intervention trajectory-aware without requiring repeated counter-context evaluation.

B.9 Fallback Behavior

ACE returns to untouched standard decoding under either of two conditions:

1. no counter-context candidate satisfies Eq. 29; or
2. the maximum layer-level reliable-evidence score $\max_{\ell<L} C^E_\ell$ is numerically negligible.

Table 5: Reference implementation profiles. Commit links identify the public source snapshots; they do not denote a tagged upstream release supporting all four backbones. Layer indices retain the source implementation's indexing convention.

Method	Source revision	Algorithm profile
DeCo	`c1a9129`	Correction strength 0.6, candidate cumulative-probability threshold 0.9, candidate cap 20, candidate layer indices 20–28.
CausalLens	`7d93bf2`	$\lambda_{\text{causal}} = 0.15$, layers 10–20 inclusive; retain the source adapter's dynamic mixing rule; optional noisy-image contrastive branch disabled.
HulluEdit	`6bc38b2`	Algorithm entries of `configs/pope_llava_run.yaml`: anchor 26, evidence/prior ranks $8/5$, $\kappa = 0.50$, $\lambda_{\text{prior}} = 0.22$; retain its robustness settings.
EnAR	`815f44f`	Stable Diffusion v1.5; DDIM schedule with 50 steps, inversion-depth index 15, 8 samples, and 512×512 reconstruction input; retain the source generation score and candidate-selection rule.

Fallback therefore introduces no preservation intervention into the original-image answer trajectory.

B.10 Generation and System Configuration

The reference execution configuration targets a single NVIDIA A800 GPU. The software recipe is Ubuntu 22.04, Python 3.10.14, PyTorch 2.6.0 with CUDA runtime 12.4, and Transformers 4.57.1, with the native InternVL and LLaVA adapters implementing the visual interfaces above. MLLM weights and activations use BF16; norms, cosine similarities, reliability statistics, and evidence-score reductions use FP32. Quantization and CPU offloading are disabled. Inference uses evaluation mode, `inference_mode`, and greedy decoding. The baseline-specific cache policies and EnAR's FP16 auxiliary diffusion precision are specified in Appendix B.11.

The reference attention path uses eager kernels exposing per-source attention-weighted values. A paired comparison must use this same attention backend, precision, and batching policy for ACE and Greedy; otherwise kernel changes would be mixed with intervention overhead. Native image preprocessing and answer parsing remain fixed within each paired benchmark evaluation. Cost profiles use batch size one and a separately specified sequence shape; they do not redefine native image preprocessing. Appendix G gives the timing recipe and the scope of each stage.

B.11 Baseline Implementation Details

The reference comparison specifies common-harness ports of DeCo, CausalLens, HulluEdit, and EnAR (Wang et al., 2025a; Ji et al., 2026; Lin et al., 2026; Liang et al., 2026). Table 5 pins the public source revisions and algorithm profiles. These revision identifiers specify upstream references; the adaptation rules below define the additional backbone interfaces. Scalar intervention settings are shared across backbones, without benchmark-specific optimization.

Backbone interfaces. Each port uses the evaluated backbone's native processor, tokenizer, decoder blocks, and output head. DeCo reads the corresponding intermediate states and applies the native final normalization and unembedding. Its candidate range is intersected with the available nonterminal layers, so a terminal state is not reused as an early exit. CausalLens uses the actual packed visual-position mask and prompt boundaries instead of fixed system and image lengths. Grouped-query attention follows the backbone's query-to-key/value-head assignment. The public adapter replaces its nominal mixing initialization with a ratio of language-path energy to total visual/language-path energy; that dynamic rule is retained rather than describing the effective mixing coefficient as a fixed 0.15. Its full-prefix, uncached execution is retained because this energy reduction depends on the query positions included in the forward pass. HulluEdit consumes the native decoder-width states and visual-token mask at its configured anchor; its subspace linear algebra runs in FP32 before the edited state is cast back to the model dtype.

EnAR port specification. The public snapshot supplies an InternVL implementation; Qwen and LLaVA require corresponding native interfaces. Original and reconstructed images use identical

native crop and packing transforms. For a vision encoder with a class token, the attention statistic is the head-averaged class-to-patch attention from its final attention block. Without a class token, it is mean incoming patch attention, averaged over valid query patches and heads in that block. Window attention is mapped back to native patch coordinates. Scores are normalized over valid patches within each view before taking the original-minus-reconstruction difference. Spatial grouping and interpolation follow the geometric support mapping in Appendix B.1. For local views, select the top 32 attention-difference units and 64 uncertainty units across all local units; for a separate global view, select 16 and 32, respectively. If there is no separate global view, use budgets 48 and 96 over the single packed unit pool. Budgets are clipped to the number of valid units; ties use packed indices. The intervention uses the union of the two sets. Replaced positions use the native tokenizer's pad-token embedding (EOS embedding if no pad token is defined), never an integer token ID copied between tokenizers. These rules specify the cross-architecture port explicitly.

Shared inputs and decoding. Within each backbone and benchmark, methods share original images, native preprocessing, task prompts, chat templates, answer parsing, and benchmark scoring. Method-specific auxiliary images remain part of the method rather than replacing the shared original input. Answer selection is greedy, with one beam and no token sampling, repetition penalty 1, and no additional length penalty. The output caps are 64 tokens for POPE, 128 for CHAIR and FREAK, and 512 for MM-Vet; native EOS stopping is enabled for benchmark scoring. These caps are common-harness settings and override upstream demonstration scripts. Diffusion randomness is separate from greedy answer selection: its per-image seed is the low 32 bits of SHA-256 over the UTF-8 string `enar:42:` followed by that image's canonical identifier.

Common environment and timing scope. The reference environment is the single-A800 configuration in Appendix B.10. The MLLMs use BF16 and an eager attention backend, with FP32 reductions; EnAR's auxiliary diffusion network uses FP16, with Diffusers 0.35.2, Accelerate 1.8.1, NumPy 1.26.4, and SciPy 1.15.3 for the reconstruction utilities. Request batch size is one, while EnAR retains its internal eight-sample reconstruction batch. KV caching is enabled for ACE, Greedy, DeCo, and the generation paths of HulluEdit and EnAR; CausalLens retains its uncached reference execution. Thus, caching is an explicit method-level difference, not an unreported optimization. The timing boundary in Appendix G includes all request-dependent preprocessing, reconstruction, visual encoding, subspace construction, forward branches, decoding, and serialization. In particular, EnAR's diffusion reconstruction is inside TTFT and E2E, not treated as a free cached input. Any retained auxiliary decoder branches also count toward time and peak memory. Cold loading and offline CCIB construction are excluded consistently. Fixed-output efficiency profiles remain distinct from natural-EOS benchmark runs.

C Mechanistic Evaluation Protocols

This section provides the evaluation procedures underlying the mechanistic analyses in Sec. 5. Diagnostic labels define the factual readout and AFE. The runtime ranking and anchor formulas do not take those labels as inputs.

C.1 Ablation Factual Effect and Evidence Ranking

For AFE-based analysis, each backbone contributes 300 POPE-Adversarial examples: 150 correctly predicted and 150 incorrectly predicted examples under standard decoding. The initial pool is selected before scoring. A valid Controlled ACE probe is required, leaving 284, 280, 286, and 276 cases. The common constant-score mask leaves 282, 277, 283, and 273 correlation cases. Capture retains all valid-probe cases in this realization.

We analyze eight normalized decoder depths,

$$\mathcal{R} = \{0.2, 0.3, \ldots, 0.9\}.$$

For a backbone with L decoder layers, each normalized depth is mapped to the nearest available non-terminal decoder layer. Ties are resolved toward the shallower layer, and the resulting layer indices remain fixed for all diagnostics on that backbone.

For visual-conditioned unit j at diagnostic layer ℓ, we remove its projected contribution,

$$c_{\ell,t_0,j} \leftarrow 0, \tag{32}$$

and replay the affected original-image decoder suffix. Its *ablation factual effect* is

$$\mathrm{AFE}_{\ell,t_0,j} = m^{\mathrm{raw}}_{L,t_0} - m^{\mathrm{abl}(\ell,j)}_{L,t_0}. \tag{33}$$

Positive AFE indicates that removing the unit reduces terminal factual preference, whereas negative AFE indicates that the intervention increases factual preference. AFE depends on the diagnostic factual hypothesis and is therefore strictly an evaluation quantity. At inference, the proposed runtime formulas do not read AFE or the ground-truth factual direction when computing R, S, V, ξ, ω, Q, or $\ell^\dagger$.

For each sample-layer pair, we compute Spearman rank correlation across visual-conditioned units between AFE and each candidate ranking score: mean attention, value norm, S, R, RS, RSV, and Q. Correlations are first averaged across the eight diagnostic depths within each example and then averaged across examples. This aggregation prevents examples with more valid units at an individual depth from receiving disproportionate weight.

Capture@K. Rank correlation evaluates ordering across all units. To additionally measure whether a score concentrates the most useful factual contributions near the top of its ranking, let

$$k = \left\lceil \frac{K|\mathcal{I}_V|}{100} \right\rceil$$

and let $\mathrm{TopK}(s)$ contain the k highest-ranked units under score s. Ties at the cutoff are resolved by native unit index. We define

$$\mathrm{Capture@}K = \frac{\sum_{j\in\mathrm{TopK}(s)}[\mathrm{AFE}_{\ell,t_0,j}]_+}{\sum_{j\in\mathcal{I}_V}[\mathrm{AFE}_{\ell,t_0,j}]_+}, \qquad [x]_+ = \max(x,0). \tag{34}$$

Sample-layer pairs with zero denominator are excluded from the aggregation. Because Eq. 34 combines effects obtained from independent single-unit interventions, it is a ranking diagnostic and should not be interpreted as the performance obtained by jointly ablating or retaining the selected units.

C.2 Anchor Alignment and Statistical Intervals

Anchor alignment additionally requires a valid probe and non-negligible evidence, giving 229, 227, 224, and 224 cases. A robust pre-terminal factual state must exist,

$$\max_{\ell<L} m_{\ell,t_0} \geq \delta.$$

The diagnostic factual peak is $\ell^\star$ from Eq. 16, while ACE selects the label-free anchor $\ell^\dagger$ from Eq. 9. Their normalized distance is

$$d_{\mathrm{anchor}} = \frac{|\ell^\dagger - \ell^\star|}{L}. \tag{35}$$

For each evaluated example, the random-anchor baseline draws 100 non-terminal layers uniformly and averages their normalized distances to $\ell^\star$. Aggregate confidence intervals for ACE anchor alignment are obtained using 10,000 sample-level bootstrap resamples.

For Fig. 3(b), a sample is marked positive when it contains at least one robust dominance reversal after weak-band states are removed. Correct and incorrect raw-model predictions are evaluated separately, and error bars denote 95% Wilson score intervals for the corresponding binomial proportions. The prevalence comparison is conditional on raw correctness and does not identify a causal effect of reversal on error. The separate analysis in Appendix H.1 uses the exact strict-silencing cohort for deletion effects and preservation outcomes, rather than identifying the mixed diagnostic sample with that cohort.

C.3 PRESERVATION-STRATEGY ABLATIONS

The intervention variants reported in Table 3(B) use the same fixed set of 180 raw Qwen3-VL strict-silencing errors and the same label-free evidence anchor $\ell^\dagger$. They differ only in where and how the validated reliability signal is applied.

Final-layer only. Selective preservation is applied only at the final decoder block:

$$\widetilde{c}_{L,t,j} = (1 + \alpha\omega_{L,j})\, c_{L,t,j}. \tag{36}$$

Anchor only. Selective preservation is applied only at $\ell = \ell^\dagger$ and is disabled at all subsequent layers. This variant tests whether locating a useful layer is sufficient without sustained downstream preservation.

Uniform suffix. To retain suffix-level intervention while removing unit selectivity, we define

$$\overline{\omega}_\ell = \frac{1}{|\mathcal{I}_V|} \sum_{j \in \mathcal{I}_V} \omega_{\ell,j}, \qquad \widetilde{c}_{\ell,t,j} = (1 + \alpha\overline{\omega}_\ell)\, c_{\ell,t,j}, \quad \ell \geq \ell^\dagger. \tag{37}$$

This variant preserves the mean layer-wise amplification strength of ACE but assigns it uniformly to visual-conditioned units.

Anchor-frozen gate. To retain unit selectivity while removing layer-specific reliability adaptation, we freeze the gate selected at the anchor:

$$\omega_j^{\text{frozen}} = \omega_{\ell^\dagger,j}, \qquad \widetilde{c}_{\ell,t,j} = \left(1 + \alpha\omega_j^{\text{frozen}}\right) c_{\ell,t,j}, \quad \ell \geq \ell^\dagger. \tag{38}$$

The current contribution $c_{\ell,t,j}$ is still recomputed from the evolving original-image trajectory. Thus, the anchor-frozen-gate variant removes only the layer dependence of the reliability estimate and does not freeze the underlying visual contribution.

Full ACE. Full ACE uses the layer-specific $\omega_{\ell,j}$ throughout the decoder suffix beginning at $\ell^\dagger$, as defined in Eq. 10.

C.4 STRICT RESCUE AND TERMINAL-MARGIN GAIN

The strict-silencing set is determined once from the unmodified model before any intervention is applied. For sample i in this fixed set, an intervention is counted as a *strict rescue* only when both conditions hold:

$$\widehat{y}_i^{\text{int}} = y_i^{\text{GT}}, \qquad m_{L,t_0}^{\text{int},i} \geq \delta. \tag{39}$$

The first condition requires recovery of the benchmark prediction, while the second requires recovery of robust factual dominance under the internal diagnostic margin.

For sample i, terminal-margin gain is

$$\Delta m_L^i = m_{L,t_0}^{\text{int},i} - m_{L,t_0}^{\text{raw},i}. \tag{40}$$

Reported $\overline{\Delta m}_L$ is the sample mean over the corresponding fixed raw strict-silencing set.

Strict rescue and aggregate benchmark Accuracy measure different quantities. The former is evaluated on a subset of raw errors, whereas Accuracy is a net statistic over the entire benchmark. Consequently, the number of strict rescues need not equal the net increase in correct predictions.

C.5 BENCHMARK AGGREGATION

POPE Accuracy is $(\text{TP} + \text{TN})/N$, and positive-class F1 is $2\text{TP}/(2\text{TP} + \text{FP} + \text{FN})$. Percentages are computed from unrounded aggregates and rounded only for presentation.

CHAIR. The captioning protocol pairs methods on the same 1,000 image IDs with one caption per image and shared prompting, image preprocessing, and generation limits. The CHAIR scorer maps object expressions to the MSCOCO object vocabulary using its synonym and compound-expression rules; reference captions and object annotations determine which mentions are unsupported (Rohrbach et al., 2018). Let $n_{b,r}$ and $h_{b,r}$ be the total and hallucinated object mentions extracted from caption r under method b. The aggregate rates are

$$\mathrm{CHAIR}_s(b) = \frac{\sum_{r=1}^{1000} \mathbf{1}[h_{b,r} > 0]}{1000}, \qquad \mathrm{CHAIR}_i(b) = \frac{\sum_{r=1}^{1000} h_{b,r}}{\sum_{r=1}^{1000} n_{b,r}}.$$

There is no equal-mention quota, output truncation to obtain such a quota, or averaging of per-caption instance ratios. A caption with no recognized object mention contributes zero to both instance counts and remains in the sentence denominator. The instance denominator is method-specific. Table 22 reports the counts behind all main-table entries; their mention totals range from 8,331 to 11,756. Additional comparisons report CHAIR_s only. An evaluation with zero mentions across its entire caption set has undefined CHAIR_i; none of the reported main-table aggregates has that condition.

FREAK: question and option weighting. The benchmark contains 1,799 questions: 1,000 multiple-choice and 799 free-form (Yin et al., 2026). For each multiple-choice question, option D remains fixed while the contents of A, B, and C take all six permutations. The model receives the same question and image with the corresponding option order. Its answer letter is mapped back to the semantic option before assigning one of three disjoint categories: correct, commonsense error, or other error. Predictions may change category across orders. For question i, let $c_{i,r}$ and $h_{i,r}$ indicate correct and commonsense-error outcomes in permutation r. Define

$$\bar{c}_i = \frac{1}{6}\sum_{r=1}^{6} c_{i,r}, \qquad \bar{h}_i = \frac{1}{6}\sum_{r=1}^{6} h_{i,r}.$$

Each free-form question receives one category, represented by indicators c_i and h_i. The reported aggregates are

$$\mathrm{Acc} = \frac{\sum_{i\in\mathrm{MC}} \bar{c}_i + \sum_{i\in\mathrm{FF}} c_i}{1799}, \qquad \mathrm{HalluRate} = \frac{\sum_{i\in\mathrm{MC}} \bar{h}_i + \sum_{i\in\mathrm{FF}} h_i}{1799}.$$

The 6,000 multiple-choice presentations do not receive 6,000 independent question weights. Numerators can be fractional after within-question averaging. These are average accuracies, not the fraction of questions answered correctly in every permutation; the latter is a separate multiple-choice consistency metric.

FREAK: answer parsing and free-form judge. Explicit final choices are parsed before scoring. A unique terminal answer letter is used directly; conflicting or absent choices invoke a semantic option parser using the same fixed judge snapshot and a separate extraction instruction. That instruction receives the output and permuted options but no correct-answer labels. It may return A–D only when the response unambiguously selects or describes that option; otherwise it returns `INVALID`. Invalid selections count as other errors and are not dropped from denominators.

The reference free-form judging configuration specifies `gpt-5-mini-2025-08-07`, with the input image, question, reference answer, designated commonsense answer, and candidate response. Model and method names are withheld. The rubric in Appendix H.4 uses the benchmark's three-category convention; it is a specified evaluation prompt rather than an assumption that any generic correctness judge implements the same rule. The API recipe uses the Responses interface, `reasoning.effort=low`, and a 4,096-token output cap; temperature, top-p, and seed are omitted. Each item starts a fresh request without tools or prior conversation. One valid judgment is retained per free-form response. An API error, incomplete response, or unparseable category triggers at most one retry with the same payload; a persistent failure is recorded separately and scored as other error. No failed item is silently removed, and no best-of-several judgment is selected. Snapshot, prompt, request settings, raw outputs, retry status, and parsed categories belong to the evaluation record.

MM-Vet. MM-Vet entries average five grades per question over 218 questions and scale the result to 100. Its denominator is independent of POPE, captioning, and FREAK, and is unchanged by the permutation convention.

Table 6: **Complete POPE results.** Entries are Accuracy/F1. Adversarial rows reuse Table 1; each split contains 3,000 cases per configuration.

Configuration	Method	Random	Popular	Adversarial
Qwen3-VL-8B	Greedy	92.33 / 92.18	90.27 / 90.03	87.93 / 87.61
Qwen3-VL-8B	ACE	93.83 / 93.71	92.17 / 91.99	94.30 / 94.21
InternVL3.5-8B	Greedy	91.80 / 91.65	89.57 / 89.33	87.03 / 86.72
InternVL3.5-8B	ACE	93.17 / 93.04	91.33 / 91.16	94.30 / 94.22
Qwen2.5-VL-7B	Greedy	90.67 / 90.48	88.20 / 87.95	85.43 / 85.12
Qwen2.5-VL-7B	ACE	92.80 / 92.66	90.73 / 90.52	94.40 / 94.35
LLaVA-OV-7B	Greedy	89.50 / 89.30	86.93 / 86.67	84.17 / 83.88
LLaVA-OV-7B	ACE	91.30 / 91.13	89.33 / 89.10	93.03 / 92.97

C.6 Interpretation of Score and Strength Ablations

The runtime gate is nonnegative, but its contribution to the factual margin need not be positive. In a local linear readout with direction w, rescaling one contribution c changes its margin contribution by $\alpha\omega w^\top c$. This quantity is negative when $w^\top c < 0$, even if $\omega > 0$. Downstream nonlinearities provide further reasons not to interpret the preservation rule as a monotonicity guarantee. Likewise, deleting a multiplicative factor in $[0, 1]$ can increase the gate magnitude. A component ablation without magnitude matching changes both selectivity and strength; a ranking diagnostic and a rescue test therefore answer different questions.

D Additional Results and Accounting

This appendix gives complete comparison tables, sensitivity analyses, fallback partitions, and paired-transition accounting.

D.1 Complete POPE Evaluation

The Random/Popular and shared Adversarial results are shown in Table 6.

D.2 Paired Main-Baseline Comparisons

We compare ACE and EnAR on the same 3,000 POPE-Adversarial example IDs for each backbone. For example i, let z_i^{ACE} and z_i^{EnAR} indicate answer correctness. The paired accuracy contrast, in percentage points, is

$$\widehat{\Delta} = \frac{100}{N} \sum_{i=1}^{N} \left(z_i^{\mathrm{ACE}} - z_i^{\mathrm{EnAR}} \right).$$

Each bootstrap replicate draws N example indices with replacement and applies the same indices to both methods. We use 20,000 replicates and report percentile intervals at the 2.5th and 97.5th percentiles, with linear quantile interpolation. Randomness uses NumPy PCG64 with `SeedSequence([20260926, b])`, where $b = 0, 1, 2, 3$ follows the backbone order in Table 7. Case IDs are sorted before resampling. Predictions are held fixed; this procedure does not repeat model inference or measure run-to-run variability.

The resampling unit is the paired example, not the image. Accordingly, these intervals quantify example-level resampling uncertainty and do not account for dependence among questions sharing an image. They are marginal 95% intervals for each contrast, not simultaneous intervals across the four backbones. All contrasts are computed from integer correctness counts before rounding.

For LLaVA-OneVision, the exact contrast is $197/3000 \times 100 = 6.566\ldots$ percentage points, reported as 6.57. Subtracting the separately rounded accuracy entries ($93.03 - 86.47$) instead gives 6.56; the paired contrast uses the unrounded counts.

Table 7: Paired ACE–EnAR accuracy contrasts on POPE-Adversarial. ACE-only and EnAR-only count examples answered correctly by only that method. Contrasts and percentile intervals are in percentage points; $N = 3{,}000$ for each backbone.

Backbone	ACE-only	EnAR-only	Δ (pp)	95% CI (pp)
Qwen3-VL-8B	295	151	+4.80	[3.43, 6.20]
InternVL3.5-8B	325	160	+5.50	[4.10, 6.90]
Qwen2.5-VL-7B	334	145	+6.30	[4.90, 7.70]
LLaVA-OneVision-7B	379	182	+6.57	[5.07, 8.10]

Table 8: **Additional comparisons.** ACE entries repeat the corresponding main-table values. CHAIR_s retains 1,000 captions. FREAK uses the question-weighted six-permutation protocol in Appendix C.5.

Configuration	Method	POPE-A Acc.	CHAIR_s	FREAK Acc.
Qwen3-VL-8B	VCD	88.40	21.7	45.3
Qwen3-VL-8B	OPERA	88.63	21.3	45.4
Qwen3-VL-8B	PTI	89.07	20.5	45.5
Qwen3-VL-8B	LTS-FS	89.23	20.1	46.1
Qwen3-VL-8B	ACE	94.30	18.9	47.8
InternVL3.5-8B	VCD	87.57	22.9	42.9
InternVL3.5-8B	OPERA	87.87	22.5	43.2
InternVL3.5-8B	PTI	88.20	21.9	43.4
InternVL3.5-8B	LTS-FS	88.43	21.5	43.7
InternVL3.5-8B	ACE	94.30	20.2	45.4
Qwen2.5-VL-7B	VCD	86.13	29.2	35.8
Qwen2.5-VL-7B	OPERA	86.47	28.8	35.8
Qwen2.5-VL-7B	PTI	87.37	27.6	36.5
Qwen2.5-VL-7B	LTS-FS	87.60	27.1	36.9
Qwen2.5-VL-7B	ACE	94.40	25.6	38.6
LLaVA-OV-7B	VCD	84.93	27.3	34.4
LLaVA-OV-7B	OPERA	85.23	26.8	34.6
LLaVA-OV-7B	PTI	85.73	25.8	34.8
LLaVA-OV-7B	LTS-FS	86.00	25.1	35.2
LLaVA-OV-7B	ACE	93.03	23.8	36.8

D.3 Additional Comparisons

Table 8 gives additional method comparisons; repeated ACE entries reuse the main table.

D.4 Reliability-Component Ablation

Table 9 compares gate allocation at Full ACE's $\alpha_0 = 0.35$ reference budget. All variants share the probe, anchor, fallback decisions, and intervention range. Variant-specific coefficients are calibrated on 1,000 separate cases using the reference-state residual injection norm, rather than answer correctness or rescue targets. Appendix F specifies the calibration and gives the second reference-budget level.

At the default budget, Full ACE gives 94.30% accuracy and 52 strict rescues, compared with 93.30%/41 without S, 94.07%/38 without V, and 94.13%/50 without $(1 - \xi)$. Uniform allocation gives 92.27% and 26 rescues; Shuffled gives 92.25% and 22.4 rescues on average. The matched-budget comparisons consistently favor structured reliability allocation over uniform and shuffled controls. Component-specific contrasts further characterize the contributions of stability, visual specificity, and edit exposure, with paired uncertainty reported in Appendix F.

R affects anchor localization but does not enter the Preserve gate. It is therefore analyzed with the original fixed-strength configurations in Table 17, where the anchor is recomputed. That table also retains the higher outcomes obtained by removing bounded gate factors without budget matching. Those rows and Table 9 use different controls and must not be interpreted as repeated estimates of the same ablation setting.

Table 9: **Component and allocation controls at a matched reference budget.** The target budget is Full ACE at $\alpha_0 = 0.35$. Each variant uses its own α_v, calibrated on 1,000 disjoint cases. All rows share Full ACE's probe, anchor, fallback decisions, 3,000 evaluation cases, and 180-case strict-silencing subset. Shuffled entries give mean $\pm$ sample standard deviation over five permutations. Coefficients are rounded for display. Test and realized budget mismatches and paired uncertainty are reported in Table 16.

Variant	α_v	POPE-A Acc.	F1	Strict rescue
Full ACE	0.350	94.30	94.21	52/180 (28.9%)
w/o S	0.232	93.30	93.20	41/180 (22.8%)
w/o V	0.182	94.07	93.97	38/180 (21.1%)
w/o ξ	0.259	94.13	94.04	50/180 (27.8%)
Uniform	0.412	92.27	92.12	26/180 (14.4%)
Shuffled	0.407 ± 0.002	92.25 ± 0.07	92.10 ± 0.07	22.4 ± 3.4/180

Table 10: **Preservation-strength sensitivity.** Features, candidates, gates, and anchors are fixed; only α changes. The zero row recovers Raw and the default row matches the main table.

α	POPE-A Acc.	F1	Strict rescue
0.00	87.93	87.61	0/180 (0.0%)
0.15	91.17	90.96	3/180 (1.7%)
0.25	92.63	92.50	27/180 (15.0%)
0.35	94.30	94.21	52/180 (28.9%)
0.45	96.00	95.95	71/180 (39.4%)
0.60	97.10	97.08	107/180 (59.4%)

Table 11: **Full-cohort acceptance and fallback.** The first three columns partition 3,000 cases. Probe-statistic means condition on validity and use a different pool from Table 3(A).

Configuration	Probe fallback	Evidence fallback	Intervention	P_{pres}	D_C
Qwen3-VL-8B	4.9%	0.7%	94.4%	0.919	0.441
InternVL3.5-8B	5.0%	0.9%	94.1%	0.919	0.441
Qwen2.5-VL-7B	6.0%	0.9%	93.1%	0.919	0.442
LLaVA-OV-7B	5.2%	0.6%	94.2%	0.919	0.442

D.5 Sensitivity to Preservation Strength

Increasing the preservation strength within the fixed Full ACE allocation rule raises POPE-Adversarial accuracy from 94.30% at $\alpha = 0.35$ to 97.10% at $\alpha = 0.60$. The main experiments use $\alpha = 0.35$ as a shared reference configuration across backbones and benchmarks.

Every row retains Full ACE's allocation rule. This sweep addresses strength sensitivity within that rule; it does not compare reliable, uniform, and shuffled allocation at equal budgets. Appendix F uses both 0.35 and 0.60 as reference levels for that comparison. A higher coefficient alone cannot establish either the necessity or the irrelevance of reliability-based weighting.

D.6 Counter-Context Acceptance and Fallback

The probe-failure, evidence-failure, and intervention counts are disjoint and exhaustive over each 3,000-case cohort. Evidence failure is counted only after a valid probe. Mean probe metrics condition on probe validity.

D.7 Inference Efficiency

The timing ledger contains 128 paired requests per method and profile at batch size one. The fixed-shape workload has 128 text positions, 576 visual positions, and exactly 64 output tokens, with no

Table 12: **Cost profiles with explicit denominators.** Decode ms/token divides total decode time by 128×63. Memory is the mean of per-request peaks. E2E and slowdown reuse Table 2.

Profile	Method	E2E (s)	TTFT (s)	Decode ms/token	Mean peak GB	Slowdown
qwen3	Greedy	2.18	0.57	25.1	18.7	1.00×
qwen3	DeCo	2.51	0.69	28.3	19.6	1.15×
qwen3	CausalLens	2.39	0.68	26.6	19.3	1.10×
qwen3	HulluEdit	2.44	0.70	27.1	19.8	1.12×
qwen3	EnAR	4.08	1.21	44.7	25.9	1.87×
qwen3	ACE	2.93	0.95	31.0	22.4	1.34×
llava	Greedy	1.96	0.50	22.8	17.4	1.00×
llava	DeCo	2.23	0.60	25.5	18.2	1.14×
llava	CausalLens	2.12	0.57	24.2	17.9	1.08×
llava	HulluEdit	2.18	0.59	24.8	18.4	1.11×
llava	EnAR	3.61	1.07	39.6	23.8	1.84×
llava	ACE	2.59	0.80	28.1	20.6	1.32×

Table 13: **Strict-silencing recovery across backbones.** Each row uses the complete fixed raw strict-silencing cohort, including fallback cases. Mean terminal-margin gains are averaged over the same complete cohorts.

Configuration	Silenced	Rescued	Rescue	$\overline{\Delta m}_L$
Qwen3-VL-8B	180	52	28.9%	1.42
InternVL3.5-8B	187	75	40.1%	1.31
Qwen2.5-VL-7B	225	94	41.8%	1.53
LLaVA-OV-7B	251	80	31.9%	1.26

early EOS. The 576-position setting defines this cost workload; it is not the native image-token count for every backbone. Native benchmark preprocessing follows Appendix B.1.

TTFT covers request preprocessing through selection of the first output token. E2E adds the remaining 63 decode steps and response serialization. Stages are sequential and non-overlapping. Slowdown is a ratio of mean E2E costs, not a mean of per-request ratios. Decode ms/token divides the total post-first-token decode cost by 128×63. Memory is the mean of per-request peak allocated GPU memory, including resident model weights; it is not peak reserved memory or the maximum across requests.

Candidate checks, selected-probe processing, the three null visual passes, reliability scoring, and suffix replay all belong to TTFT. Probe fallback retains completed search cost; evidence fallback retains completed validation cost; neither incurs suffix replay. Offline bank construction, cold model loading, warmup, and network transfer are excluded. Appendix G gives the synchronization protocol, short-output comparison, and explicit stage accounting.

D.8 Strict Rescue and Aggregate Accuracy

We distinguish strict-silencing recovery from aggregate accuracy improvement. Strict recovery is evaluated on the fixed raw strict-silencing cohort of each backbone, whereas aggregate accuracy uses all 3,000 paired cases. Table 13 reports the strict-cohort recovery results; the subsequent transition accounting explains how these recoveries contribute to overall accuracy.

All transition counts use paired Raw and ACE predictions for the same 3,000 cases per configuration. Correctness is evaluated against the same reference answer under the same scoring rule. We count E→C and C→E directly from these pairs rather than inferring either count from aggregate accuracy. The net change in correct predictions is

$$G = N_{\text{corr}}^{\text{ACE}} - N_{\text{corr}}^{\text{raw}} = N_{\text{E}\to\text{C}} - N_{\text{C}\to\text{E}}. \tag{41}$$

Let $\mathcal{S}$ be the fixed Raw strict-silencing error cohort. The E→C transitions partition into three disjoint categories:

Table 14: **Paired-transition accounting and correction decomposition.** Panel A uses all 3,000 paired cases per configuration; Raw C and ACE C denote correct-prediction counts, and Net is E→C minus C→E. Panel B partitions every E→C transition into strict rescue (SR), other strict-cohort correction (SC), or non-strict-cohort correction (NC). The cohort is determined from Raw and held fixed. Counts are not averages across repeated runs.

A. Full-cohort transitions

Configuration	Raw C	ACE C	E→C	C→E	Net
Qwen3-VL-8B	2638	2829	191	0	191
InternVL3.5-8B	2611	2829	218	0	218
Qwen2.5-VL-7B	2563	2832	269	0	269
LLaVA-OV-7B	2525	2791	266	0	266

B. Disjoint decomposition of corrections

Configuration	SR	SC	NC	E→C
Qwen3-VL-8B	52	19	120	191
InternVL3.5-8B	75	19	124	218
Qwen2.5-VL-7B	94	30	145	269
LLaVA-OV-7B	80	45	141	266

- **Strict rescue (SR):** a case in $\mathcal{S}$ becomes correct and satisfies $m^{\mathrm{ACE}}_{L,t_0} \geq \delta$.
- **Other strict-cohort correction (SC):** a case in $\mathcal{S}$ becomes correct but satisfies $m^{\mathrm{ACE}}_{L,t_0} < \delta$.
- **Non-strict-cohort correction (NC):** a Raw error outside $\mathcal{S}$ becomes correct, irrespective of its ACE terminal margin.

Consequently,

$$N_{\mathrm{E}\to\mathrm{C}} = N_{\mathrm{SR}} + N_{\mathrm{SC}} + N_{\mathrm{NC}}, \qquad N_{\mathrm{other}} = N_{\mathrm{SC}} + N_{\mathrm{NC}}.$$

The transition identity also gives

$$N_{\mathrm{C}\to\mathrm{E}} = N_{\mathrm{SR}} + N_{\mathrm{SC}} + N_{\mathrm{NC}} - G \geq \max(0, N_{\mathrm{SR}} - G). \tag{42}$$

The lower bound alone does not determine the regression count; that count requires the paired predictions.

Strict recovery versus aggregate improvement. Table 13 reports strict-rescue rates of 28.9–41.8% across the four backbones. These rates use backbone-specific strict-silencing cohorts rather than the complete 3,000-case evaluation sets.

For Qwen3, Table 14 gives $G = 2829 - 2638 = 52 + 19 + 120 - 0 = 191$. Within its 180-case strict-silencing cohort, 71 cases become correct, including 52 strict rescues and 19 corrections below the robust terminal-margin threshold. The corresponding answer-correction and strict-rescue rates are $71/180 = 39.4\%$ and $52/180 = 28.9\%$.

Strict silencing is a post hoc diagnostic criterion, whereas runtime intervention eligibility depends on probe validity and available evidence. ACE can therefore correct cases outside the strict-silencing cohort. Such corrections contribute to aggregate accuracy but cannot be attributed to recovery from the defined strict-silencing failure mode solely from their corrected answers.

Paired Prediction Stability and Terminal-Margin Dynamics. All four paired cohorts have zero C→E transitions. Nevertheless, 2, 3, 0, and 3 initially correct cases, respectively in the configuration order of Table 14, have lower terminal factual margins under ACE and remain correct. Thus, unchanged answer correctness does not imply nondecreasing margin. The gate is nonnegative, but the contribution it amplifies may have either factual direction, and subsequent decoder transformations can further change the effect. Neither zero regressions in these cohorts nor the form of the intervention establishes a general no-regression guarantee.

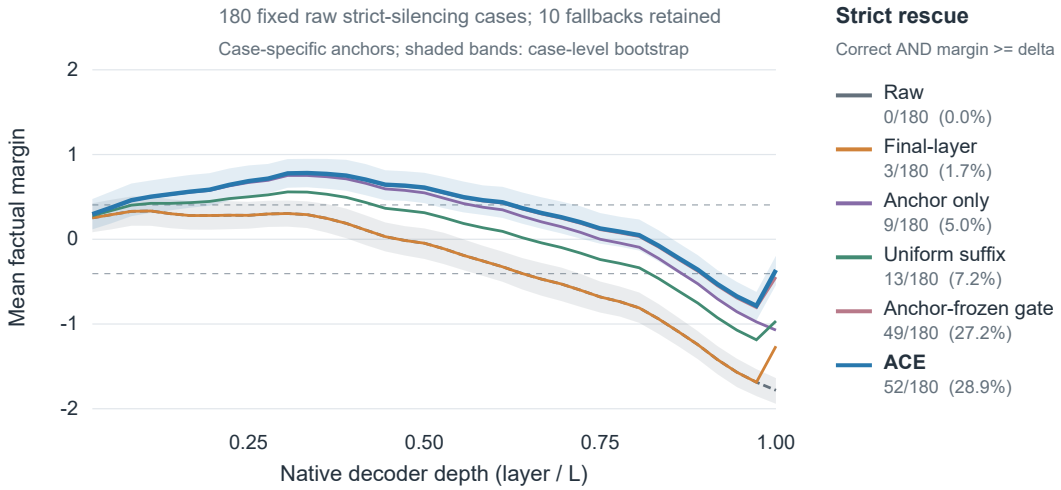


Figure 6: **Preservation trajectories for the fixed Qwen3 cohort.** All six curves average the same 180 frozen strict-silencing cases, including ten fallback cases. The anchor-frozen variant starts at each case's selected anchor, freezes only its gate, and recomputes current contributions; it is not a fixed-mid-layer intervention. The right panel reports correct-and-robust strict rescues from these same case outputs. Dashed horizontal lines mark $\pm\log(1.5)$. Bands for Raw and ACE are 95% pointwise bootstrap intervals from 3,000 resamples of the cohort.

E Trajectory Analysis and Qualitative Examples

E.1 Layer-wise Preservation Trajectories

Figure 6 uses the same 180 cases and per-case intervention outputs as Table 3(B). The horizontal axis is native normalized decoder depth, ℓ/L; it is not a common post-anchor coordinate. Case-specific intervention onsets therefore occur at different horizontal positions.

Terminal means and strict-rescue percentages are separate summaries. In particular, ACE's mean terminal margin remains negative while 52 individual cases meet the positive robust-margin and correctness conditions. The mean and the thresholded case count summarize different features of the terminal-margin distribution and need not have the same sign.

E.2 Counter-Context Examples

Figure 7 illustrates the ACE workflow through three admissible-probe examples and one fallback example. These examples are independent of the quantitative cohorts and do not contribute to the reported accuracy, strict-rescue, or fallback denominators. In particular, the counting question is not a POPE binary decision.

The first three examples depict query-conditioned placement followed by validation, anchor localization, and preservation. An answer correction alone is insufficient to establish strict rescue, which additionally requires the terminal-margin criterion. In the fourth example, the displayed candidate has $P_{\text{pres}} = 0.79 < 0.90$ and $D_C = 0.67 > 0.60$, violating both constraints. Fallback represents failure of all evaluated candidates, not rejection of this candidate alone. No evidence anchor is selected and standard decoding is retained.

E.3 Fallback and Failure Modes

Probe fallback requires no admissible candidate; evidence fallback requires a valid probe but negligible evidence. Both retain standard decoding, with different accumulated costs. Failed rescues remain in the denominator; corrections below the margin threshold count only toward E→C.

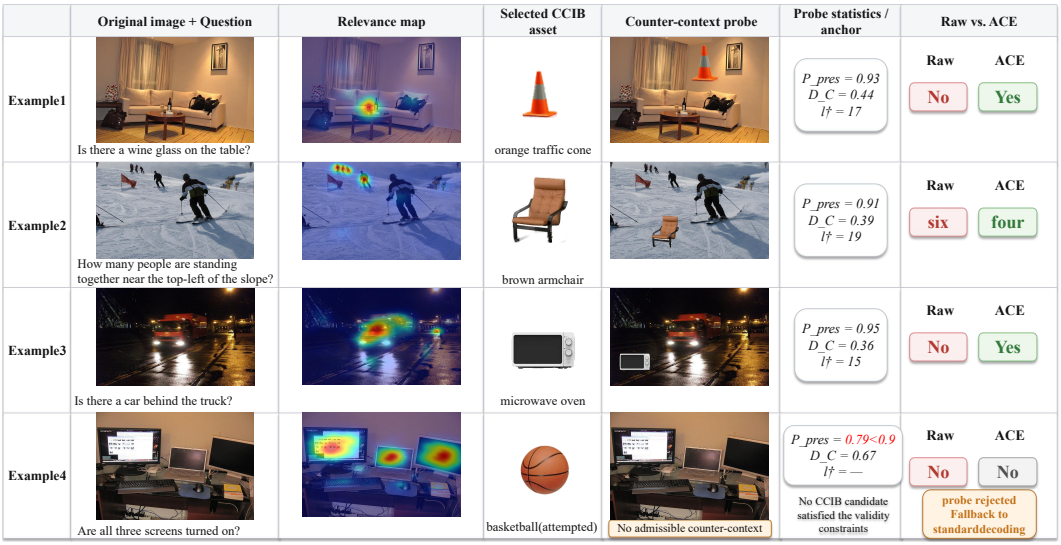


Figure 7: **Counter-context examples.** Columns show the image/question, relevance map, selected or attempted CCIB asset, counter-context probe, probe statistics and anchor, and Raw/ACE answers. The first three examples depict admissible probes. The fourth depicts fallback when no candidate is admissible; the shown candidate violates both validity constraints. These workflow examples are independent of the quantitative cohorts.

F Gate Allocation at Matched Reference Budgets

F.1 Shared Cases and Controlled Variants

The comparison uses the same 3,000 Qwen3 cases and fixed 180-case Raw strict-silencing subset as the main analysis. Every variant inherits Full ACE's selected probe, anchor $\ell_i^\dagger$, intervention eligibility $a_i \in \{0, 1\}$, and decoder suffix. Fallback cases remain unchanged and stay in the denominators. The comparison isolates changes to gate allocation under fixed intervention support; it is distinct from the configuration-level removals in Table 17.

With case indices suppressed, the six gates are

$$g_{\ell,j}^{\text{Full}} = (1-\xi_j)S_{\ell,j}V_j, \qquad g_{\ell,j}^{-S} = (1-\xi_j)V_j,$$
$$g_{\ell,j}^{-V} = (1-\xi_j)S_{\ell,j}, \qquad g_{\ell,j}^{-\xi} = S_{\ell,j}V_j,$$
$$g_{\ell,j}^{\text{Uniform}} = \frac{1}{|\mathcal{I}_V|}\sum_{k\in\mathcal{I}_V} g_{\ell,k}^{\text{Full}}, \quad g_{\ell,j}^{\text{Shuffled}} = g_{\ell,\pi_\ell(j)}^{\text{Full}}.$$

Permutations act within each case and layer's eligible visual units and remain fixed throughout replay. Five realizations use seeds 17, 29, 43, 71, and 101; each receives a separate amplitude calibration. Calibration and evaluation use separate permutation streams. Means and sample standard deviations across these five realizations are reported for Shuffled. The other variants each have one deterministic result.

R enters $Q = R\omega$ for anchor localization, not the Preserve gate. With eligibility, anchor, and gate fixed, removing R would not change the intervention; its localization effect remains in Table 17 rather than this gate-allocation test.

F.2 Reference Budget and Independent Calibration

A separate 1,000-case calibration set determines the amplitude coefficients before test scoring. The calibration-source protocol draws 1,000 distinct images from MS-COCO 2014 `train2014`, using `instances_train2014.json` to construct one object-existence question per image. This is a constructed calibration set, not an official POPE training split. Candidate images with a canonical

image ID or decoded-RGB SHA-256 matching any evaluation image are removed before selection. Decoded-image hashes include width, height, and RGB bytes, and detect exact duplicates rather than all near-duplicates.

Eligible images are sorted by SHA-256 of the UTF-8 string `ace-cal-v1:20260926:` followed by the zero-padded 12-digit COCO image ID; ties use image ID. The first 1,000 images are selected. In this order, alternating images receive a present or absent category, giving 500 questions of each type. Within the appropriate category set, the category with smallest SHA-256 of `ace-cal-object-v1:20260926:`*image-ID*`:`*category-ID* is selected; IDs use 12 and 3 digits, respectively, and ties use category ID. The prompt is `Is there a <category> in the image?`, using the COCO category name verbatim. Images without an eligible present or absent category are excluded before sorting.

Annotations only define the calibration queries. Answer labels, prediction correctness, rescue outcomes, and AFE do not enter amplitude fitting. These cases are used only for the variant-specific budget coefficients, not to choose ACE's default thresholds, reference strengths, CCIB assets, or baseline parameters. Calibration cases are image-disjoint from the evaluation set by the ID and exact-image checks above; this statement does not assert disjointness from model pretraining. With reference contributions expressed in the common residual space, define

$$d_{i,\ell}^{(v)}(\alpha) = a_i \mathbf{1}\{\ell \geq \ell_i^{\dagger}\}\alpha \sum_{j \in \mathcal{I}_V} g_{i,\ell,j}^{(v)} c_{i,\ell,t_0,j}^{\mathrm{ref}},$$

$$B_v(\alpha; \mathcal{C}) = \frac{1}{|\mathcal{C}|} \sum_{i \in \mathcal{C}} \sum_{\ell=1}^{L} \left\| d_{i,\ell}^{(v)}(\alpha) \right\|_2 .$$

The reference is the unmodified original-image trajectory. Fallback cases contribute zero while remaining in the average. The norm follows the sum over units and therefore accounts for within-layer cancellation; norms, rather than vectors, are summed across layers.

For each Full ACE reference $\alpha_0 \in \{0.35, 0.60\}$,

$$\alpha_v = \alpha_0 \frac{B_{\mathrm{Full}}(1; \mathcal{C})}{B_v(1; \mathcal{C})}.$$

All denominators are positive. Fixed reference states imply linear scaling of this budget with nonnegative α, yielding equal mean calibration budgets up to numerical precision. Calibration uses no answer correctness, rescue outcome, or AFE target. The fitted coefficients remain frozen on the evaluation set.

For a set $\mathcal{D}$, the signed reference-budget mismatch is

$$\epsilon_v^{\mathrm{ref}} = 100 \left[\frac{B_v(\alpha_v; \mathcal{D})}{B_{\mathrm{Full}}(\alpha_0; \mathcal{D})} - 1 \right].$$

The realized budget substitutes each variant's evolving contribution for c^{ref} at the same first-answer position t_0. Its mismatch $\epsilon_v^{\mathrm{real}}$ uses the corresponding Full ACE realized budget as denominator. Reference matching neither enforces equal per-case budgets nor guarantees equal realized injections. Table 16 reports both mean mismatches; per-case budgets are retained in the accompanying records.

F.3 Outcomes and Weight Alignment

Accuracy and F1 use all 3,000 paired cases. Strict rescue uses the same 180-case denominator, and C→E counts compare each variant's answers with Raw. Table 15 shows the two budget levels; its lower-budget panel repeats Table 9. At the lower budget, Full ACE exceeds Uniform and Shuffled in accuracy by 2.03 and 2.05 percentage points, respectively; the higher-budget differences are 2.37 and 2.47 points. These comparisons favor reliability-based allocation over uniform or permuted allocation under the controlled support and approximately matched realized budgets.

AFE alignment uses $R_{\ell,t_0,j} g_{\ell,j}^{(v)}$, shared Raw-trajectory AFE targets, and the original common exclusion mask: 282 cases and 2,256 case–layer pairs. Correlations are averaged across eligible layers within each case and then across cases. Full ACE has alignment 0.503, compared with 0.406

Table 15: **Gate allocation at matched reference budgets.** Each panel uses the Full ACE strength α_0 to define a calibration budget; each variant uses its own calibrated α_v. Acc./F1 are percentages on 3,000 cases; SR counts strict rescues out of 180; C→E counts regressions against Raw. Shuffled entries are mean ± sample standard deviation over five permutations, so its mean counts need not be integers. Displayed coefficients are rounded; replay uses the full-precision calibrated values.

Variant	α_v	Acc.	F1	SR	C→E
Reference $\alpha_0 = 0.35$					
Full ACE	0.350	94.30	94.21	52	0
w/o S	0.232	93.30	93.20	41	1
w/o V	0.182	94.07	93.97	38	0
w/o ξ	0.259	94.13	94.04	50	0
Uniform	0.412	92.27	92.12	26	5
Shuffled	0.407 ± 0.002	92.25 ± 0.07	92.10 ± 0.07	22.4 ± 3.4	6.0 ± 2.8
Reference $\alpha_0 = 0.60$					
Full ACE	0.600	97.10	97.08	107	0
w/o S	0.398	96.47	96.44	94	1
w/o V	0.312	96.67	96.64	97	0
w/o ξ	0.444	97.00	96.98	104	0
Uniform	0.707	94.73	94.66	68	5
Shuffled	0.697 ± 0.003	94.63 ± 0.22	94.56 ± 0.23	62.2 ± 3.8	11.2 ± 1.8

Table 16: **Budget checks, alignment, and paired accuracy contrasts.** A: AFE rank alignment and signed test-budget mismatches in percent. Reference mismatch is identical across the two levels by linear scaling; realized mismatch is level-dependent. Shuffled entries give mean ± sample standard deviation over five realizations. B: Full minus variant accuracy in percentage points with conditional 95% paired bootstrap intervals from 3,000 resamples. For Shuffled, the contrast uses mean per-case correctness across its five fixed realizations; the interval does not additionally resample seeds.

A. Alignment and budget checks

Variant	AFE ρ	ϵ^{ref}	$\epsilon^{\text{real}}_{0.35}$	$\epsilon^{\text{real}}_{0.60}$
Full ACE	0.503	0.00	0.00	0.00
w/o S	0.406	-0.53	-0.58	-0.62
w/o V	0.450	0.05	0.00	-0.03
w/o ξ	0.480	-0.53	-0.55	-0.57
Uniform	0.303	-0.70	-0.83	-0.92
Shuffled	0.225 ± 0.003	-0.57 ± 0.63	-0.71 ± 0.63	-0.82 ± 0.63

B. Paired accuracy differences

Comparator	$\alpha_0 = 0.35$	$\alpha_0 = 0.60$
w/o S	1.00 [0.67, 1.37]	0.63 [0.37, 0.93]
w/o V	0.23 [0.00, 0.50]	0.43 [0.17, 0.70]
w/o ξ	0.17 [0.00, 0.37]	0.10 [-0.10, 0.30]
Uniform	2.03 [1.53, 2.53]	2.37 [1.83, 2.93]
Shuffled	2.05 [1.61, 2.49]	2.47 [2.00, 2.97]

without S, 0.450 without V, 0.480 without $(1 - \xi)$, 0.303 for Uniform, and 0.225 for Shuffled. Multiplying a score by its positive calibrated coefficient does not change its Spearman ranking. The Full and Uniform alignments reproduce the Q and R diagnostics, respectively, providing a consistency check across the analyses.

F.4 Interpretation and Scope

Across all individual runs, the largest absolute test reference-budget mismatch is 1.54%; the largest realized-budget mismatch is 1.78%. The matched comparison reverses the ordering between Full ACE and the fixed-α removal configurations. This is consistent with the importance of accounting

Table 17: **Fixed-strength configuration ablations.** Candidate search is fixed. Remaining factors determine the gate, anchor, and evidence fallback. Rows are not matched in amplification magnitude.

Variant	POPE-A Acc.	F1	Strict rescue
w/o R	92.53	92.38	26/180 (14.4%)
w/o S	94.47	94.39	56/180 (31.1%)
w/o V	96.43	96.40	95/180 (52.8%)
w/o ξ	95.70	95.64	67/180 (37.2%)
Full ACE	94.30	94.21	52/180 (28.9%)

for amplification and intervention support; it does not imply that magnitude alone explains every difference in Table 17, where the anchor also changes.

The component effects are not equally strong. The paired accuracy advantage over w/o S is positive at both budget levels. For w/o V, the lower-budget accuracy interval reaches zero, although the strict-rescue advantage is 7.78 percentage points with interval [3.33, 12.22]; the higher-budget accuracy contrast is positive. For w/o $(1 - \xi)$, the accuracy advantages are only 0.17 and 0.10 points, with intervals reaching or crossing zero. Its strict-rescue contrasts also include zero. Thus, the allocation comparison provides clearer support for the full weighting rule over unstructured controls than for a uniform, independent task benefit from every factor.

Higher strength remains beneficial within the Full allocation rule, so the comparison does not establish 0.35 as optimal. Neither a lower C→E count in this cohort nor a better AFE ranking establishes a general no-regression guarantee. The matched endpoints concern binary answer accuracy and first-answer margins; they do not establish a CHAIR or FREAK tradeoff. Those benchmarks require outputs from the corresponding gate variants under their own generation protocols.

F.5 Fixed-Strength Configuration Ablations

Table 17 retains the original configuration-level ablations at $\alpha = 0.35$. Candidate-search results are fixed, but the remaining score determines each variant's anchor and evidence fallback. Removing R changes localization from $\sum_j R_{\ell,t_0,j}\omega_{\ell,j}$ to $\sum_j \omega_{\ell,j}$ while preserving the gate. The other removals change the gate as well as localization; no perturbation-budget normalization is applied.

Removing S, V, or $(1 - \xi)$ improves the displayed outcomes in this configuration-level comparison. These results are retained because they show the importance of distinguishing fixed coefficient from fixed perturbation budget. They do not isolate weight allocation: deleting a bounded factor changes effective amplification, and the intervention onset and eligibility may also change. The matched-support comparison in Table 9 addresses a different question and does not supersede these results as a measure of the original configurations.

G Execution Protocol and Response-Length Accounting

G.1 Timing Boundaries and Paired Workloads

The execution protocol uses the single-GPU reference configuration in Appendix B.10. It specifies 20 warmup requests outside the reporting window, followed by 128 paired requests per method, profile, and output length. These are request averages, not averages over repeated benchmark runs. Input shape and request IDs are shared across methods. Output lengths $T \in \{1, 8, 64\}$ are fixed, with early EOS disabled for this comparison. The short settings isolate response-length dependence; they do not estimate the natural output length distribution of POPE.

The wall-clock timer is `time.perf_counter_ns`, with device synchronization before request entry and at the first-token and final output boundaries. TTFT ends when the first output token is available; E2E ends after local response serialization. GPU-stage diagnostics use CUDA events on the same stream, synchronized before reading elapsed times. CPU composition, lookup, and serialization use wall-clock intervals. For additive reporting, the ledger uses disjoint synchronized wall-clock stage intervals, including CPU launch and wait time; nested CUDA event durations are

Table 18: **Response-length cost profiles.** Means over 128 paired requests per configuration. All times are seconds; ρ_T is mean E2E divided by the matched Greedy mean at length T. The 64-token columns reproduce Tables 2 and 12. TTFT is shared across the three controlled lengths.

Method	TTFT	E2E_1	ρ_1	E2E_8	ρ_8	E2E_{64}	ρ_{64}
Qwen3 profile							
Greedy	0.569	0.599	1.00×	0.765	1.00×	2.180	1.00×
DeCo	0.694	0.727	1.21×	0.914	1.20×	2.510	1.15×
CausalLens	0.682	0.714	1.19×	0.890	1.16×	2.390	1.10×
HulluEdit	0.700	0.733	1.22×	0.912	1.19×	2.440	1.12×
EnAR	1.207	1.264	2.11×	1.560	2.04×	4.080	1.87×
ACE	0.948	0.977	1.63×	1.182	1.55×	2.930	1.34×
LLaVA profile							
Greedy	0.497	0.524	1.00×	0.675	1.00×	1.960	1.00×
DeCo	0.595	0.624	1.19×	0.792	1.17×	2.230	1.14×
CausalLens	0.569	0.595	1.14×	0.755	1.12×	2.120	1.08×
HulluEdit	0.590	0.618	1.18×	0.782	1.16×	2.180	1.11×
EnAR	1.065	1.115	2.13×	1.377	2.04×	3.610	1.84×
ACE	0.795	0.820	1.57×	1.006	1.49×	2.590	1.32×

not added a second time. No inter-request overlap is assumed. Peak allocated memory is reset after warmup and before each request and summarized over the same request set.

For request r, the accounting identity is

$$\text{E2E}_r(T) = \text{TTFT}_r + \sum_{k=2}^{T} d_{r,k} + s_r,$$

where $d_{r,k}$ is the cost of producing post-first token k and s_r is response serialization. For $T = 1$, the sum is empty; there is no post-first-token decode denominator. The length comparison holds each request's pre-answer work and serialization cost fixed and uses prefixes of its decode sequence. Thus, it isolates the cost of answer length within the fixed-shape profile rather than changes in probe selection.

G.2 Short-Output Cost Profiles

Table 18 reports the same method/profile pairs at three output lengths. ACE's TTFT is 0.948 seconds in the Qwen3 profile and 0.795 seconds in the LLaVA profile. Its one-token slowdowns are $1.63\times$ and $1.57\times$, compared with $1.34\times$ and $1.32\times$ at 64 tokens. Pre-answer work is therefore a larger fraction of short-answer cost. These ratios should be interpreted with their output-length and sequence-shape conditions.

G.3 Disjoint Pre-Answer Stages

Table 19 separates the operations that precede the first answer. Original prefill includes the original native vision frontend, language decoder, and source-contribution collection. The selected-probe prefill retains layer-wise values for stability but does not repeat the original stream's source-contribution collection. Search comprises bank lookup, one cached original-image validity encoding, and candidate composition/validity passes. The selected probe has a native visual pass and one language prefill. The three null references each have their own native visual pass and no language prefill. Reliability scoring includes feature comparisons, support exposure, score aggregation, and anchor selection; suffix replay then operates only on the original-image stream.

Entries average over all 128 requests, including zero costs for stages skipped by fallback. They are not conditional means over accepted probes. The candidate budget is eight. Additional native visual passes are charged to TTFT regardless of whether they invoke a language decoder. The 64-token E2E slowdown alone cannot identify their individual cost.

Table 19: **ACE stage accounting.** Mean milliseconds over the same 128 requests per profile. Stages above TTFT are disjoint; null references are processed sequentially. Totals use unrounded values. The final three rows recover the 64-token E2E costs.

Stage	Qwen3 profile	LLaVA profile
Request preprocessing	24.43	20.49
Original native vision + language prefill	410.34	344.27
CCIB lookup	7.01	5.91
Original validity encoding $\Phi(I)$	10.70	8.84
Candidate composition + validity checks	70.22	59.03
Selected-probe native vision	31.43	26.47
Selected-probe language prefill	98.87	83.07
Mean-color null: native vision	31.62	26.48
Blurred null: native vision	31.24	26.40
Noise null: native vision	31.55	26.11
Reliability scoring + anchor selection	34.19	28.69
Original-stream suffix replay	127.01	106.56
First-token selection	39.08	32.79
TTFT subtotal	947.69	795.11
Remaining 63 decode steps	1953.00	1770.30
Response serialization	29.31	24.59
64-token E2E total	2930.00	2590.00

H Cohort-Aligned Mechanism and Benchmark Accounting

H.1 Functional Evidence Within the Strict-Silencing Cohort

This analysis uses the same 180 Qwen3 raw strict-silencing cases as Fig. 3(a) and Table 3(B). Case IDs, raw trajectories, reliability scores, anchors, fallback decisions, and preservation outcomes are fixed. The mixed correct/error diagnostic pool used in Fig. 4 is a different analysis population and is not substituted for these cases.

For each case, the reference peak is the earliest non-terminal layer attaining the maximum raw factual margin. Its analysis window contains that layer and its immediate neighbors, clipped to layers $1, \ldots, L-1$ and deduplicated. The window is selected before consulting AFE or ACE outcomes. At each window layer, units are ranked by Q or mean attention; the top $\lceil 0.1|\mathcal{I}_V| \rceil$ are retained, with native unit index breaking ties. Single-unit AFE uses the raw-stream deletion and suffix replay in Appendix C.1.

Three summaries are averaged over layers within each case and then over cases: mean signed AFE of selected units, their fraction with positive AFE, and their capture of the total positive AFE. Uniform random selection uses its exact conditional expectations: the mean and positive fraction over all units and capture $k/|\mathcal{I}_V|$. No selected-unit effects are summed and interpreted as a joint intervention. Top-10% selection is only a diagnostic budget: ACE itself applies continuous gates from its anchor and does not replace them with a hard top-k mask. This analysis is therefore not a mediation decomposition of the total preservation effect.

Nine probe fallbacks and one evidence fallback remain in the 180-case outcome cohort. Ranking and association summaries condition on the 170 cases with active preservation, where the gate has an operational meaning. They are not reported as unconditional 180-case averages. The positive-AFE capture denominator is nonzero in every analyzed case-layer pair. Intervals resample whole cases, retaining all their layers and paired rankings together (10,000 bootstrap replicates; seed 2026092503).

Top-Q signed AFE exceeds the attention and random expectations by 0.149 $[0.113, 0.187]$ and 0.241 $[0.193, 0.291]$, respectively. Across the 170 active cases, its Spearman association with terminal-margin gain is 0.168 $[0.012, 0.317]$. Its association with the binary rescue indicator, computed as the correlation between average AFE ranks and that indicator, is 0.069 $[-0.083, 0.224]$. The 52 rescued cases have mean selected signed AFE 0.314; the 118 active non-rescued cases have mean 0.274. These exploratory associations do not adjust for raw terminal margin, anchor depth, or other case

Table 20: **Functional effects around the raw factual peak.** The 170 active cases belong to the fixed 180-case strict cohort. Brackets give case-bootstrap 95% intervals. Positive fraction and capture are percentages. Random reports its exact selection expectation.

Ranking	Signed AFE	Positive fraction	Capture@10
Q	0.286 [0.231, 0.345]	88.1 [85.9, 90.1]	37.1 [35.8, 38.5]
Attention	0.137 [0.099, 0.180]	67.3 [64.3, 70.1]	22.2 [20.8, 23.5]
Random	0.045 [0.036, 0.055]	55.7 [54.5, 56.8]	10.0 [10.0, 10.0]

Table 21: **Paired strict-rescue outcomes.** All 180 cases, including fallbacks, are retained. Marginals reproduce 52/180 for Full ACE and 49/180 for Anchor-frozen gate in Table 3(B).

	Frozen rescued	Frozen not rescued	Total
ACE rescued	48	4	52
ACE not rescued	1	127	128
Total	49	131	180

difficulty. They provide a within-cohort connection between ranked contributions and functional effects, but do not show that positive AFE alone determines recovery. Deletion and amplification need not be inverse operations in a nonlinear model, and a prevalence comparison cannot by itself prove that reversal caused an incorrect prediction.

H.2 Paired Outcomes for Frozen and Layer-Specific Gates

Anchor-frozen and Full ACE share the candidate, onset layer, fallback decisions, preservation strength, and per-token contribution recomputation. They differ in whether the reliability gate depends on the current layer. Strict rescue requires both a correct final answer and terminal margin at least δ for the same raw strict case.

The net difference is three cases, comprising four gains and one loss relative to the frozen gate. Resampling paired outcomes yields a 1.67-percentage-point difference with 95% interval $[-0.56, 4.44]$. The exact two-sided McNemar test uses the five discordant pairs and gives $p = 0.375$. The continuous terminal-margin difference is 0.084 $[0.061, 0.107]$, averaged over all cases. Both intervals use 10,000 case-bootstrap replicates. Thus, the continuous endpoint improves in this comparison, while the binary rescue difference is not resolved at conventional significance levels. The larger comparison with one-shot and uniform-suffix variants supports sustained selective preservation; it should not be described as a large independent benefit from layer-specific gating.

H.3 CHAIR Numerators and Method-Specific Denominators

Table 22 gives the aggregation counts used in Table 1. Every row has 1,000 captions. H_s is the number containing a hallucinated mention, H_i the number of hallucinated mentions, and M_i the total extracted mentions. Accordingly, $\text{CHAIR}_s = 100H_s/1000$ and $\text{CHAIR}_i = 100H_i/M_i$. Changes in output object coverage affect the instance denominator; a lower sentence rate need not imply the lowest instance rate among methods.

H.4 FREAK Judge and Option-Extraction Rubrics

The free-form judge receives a fixed instruction followed by the input image and fields for the question, reference answer, commonsense answer, and response. The instruction is:

> Assign the response exactly one label: CORRECT, COMMONSENSE, or OTHER. Treat the supplied answer annotations as authoritative for this question; do not replace unusual visual facts with what normally occurs. Evaluate only the information requested. A response that agrees with the reference answer without a material contradiction is CORRECT. Otherwise, a response that commits to the designated commonsense alternative without also asserting an incompatible reference answer is COMMONSENSE. Unsupported alternatives, unresolved

Table 22: **CHAIR aggregation counts.** Method-specific M_i replaces a common mention quota. Ratios are rounded only after aggregation.

Configuration	Method	H_s	H_i	M_i
Qwen3-VL	Greedy	228	690	10877
Qwen3-VL	DeCo	219	650	8887
Qwen3-VL	CausalLens	208	610	9217
Qwen3-VL	HulluEdit	197	570	9752
Qwen3-VL	EnAR	201	590	10637
Qwen3-VL	ACE	189	530	8939
InternVL3.5	Greedy	241	740	10337
InternVL3.5	DeCo	230	700	11756
InternVL3.5	CausalLens	222	660	10239
InternVL3.5	HulluEdit	210	610	9761
InternVL3.5	EnAR	214	630	9127
InternVL3.5	ACE	202	570	9555
Qwen2.5-VL	Greedy	306	840	9534
Qwen2.5-VL	DeCo	294	800	11518
Qwen2.5-VL	CausalLens	280	740	10514
Qwen2.5-VL	HulluEdit	267	690	8428
Qwen2.5-VL	EnAR	272	700	9623
Qwen2.5-VL	ACE	256	640	9859
LLaVA-OV	Greedy	286	880	10966
LLaVA-OV	DeCo	272	830	10442
LLaVA-OV	CausalLens	259	770	10319
LLaVA-OV	HulluEdit	246	720	8331
LLaVA-OV	EnAR	251	740	10161
LLaVA-OV	ACE	238	670	10947

> contradictions, refusals, and empty answers are OTHER. Mere quotation or explicit rejection of the commonsense alternative is not commitment to it. When an annotation marks acceptable alternatives with hash delimiters, matching any listed alternative is sufficient for that annotation, subject to the same contradiction rule. Return a JSON object with keys category and reason; category must be one of the three labels and reason must be brief.

The input records retain the annotation text and any delimiters exactly. The judge configuration and retry handling are specified in Appendix C.5. All three labels are retained for main-table Accuracy and HalluRate; additional baseline entries report Accuracy only.

For an unformatted multiple-choice response, the separate extraction instruction is:

> Read the response and the displayed options. Identify the option that the response ultimately selects or unambiguously describes. Do not solve the visual question or evaluate whether the choice is correct. Return one of A, B, C, D, or INVALID. Use INVALID when no unique intended option can be extracted, including an unresolved conflict between final choices.

This parser receives no correctness or commonsense-category annotations. Scoring follows only after mapping its output through the displayed permutation. Six-presentation records retain the question ID, option mapping, extracted answer, semantic category, and per-question mean, so that permutation averaging cannot silently change the sample weight.