VERIFY BEFORE EDIT: COUNTERFACTUAL VISUAL EVIDENCE VERIFICATION FOR MLLM HALLUCINATION MITIGATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal large language models (MLLMs) often produce plausible responses unsupported by the visual input. Existing training-free mitigation methods typically suppress language priors or amplify visual representations, but rarely verify whether the targeted visual evidence is relevant to the current decision. We introduce **VBEDIT**, a training-free **"verify before edit"** framework for correcting visually unsupported predictions in binary visual queries. Before editing the first-step answer representation, VBEDIT evaluates the decision relevance of answer-conditioned visual evidence. It computes signed visual-token attributions with respect to the semantic margin between competing answer candidates, separating supporting from opposing evidence. It then counterfactually perturbs the most supportive tokens and measures the resulting margin change as a sample-specific test of decision relevance. This counterfactual response gates a contrastive edit of the answer representation, without requiring external detectors, reference responses, or auxiliary supervision. Extensive experiments across diverse MLLMs and benchmarks show that VBEDIT improves over matched Vanilla and representative methods. VBEDIT also improves accuracy and F1 on AMBER-D at both LLaVA scales and raises HOPE accuracy by 5.93 and 9.43 percentage points on LLaVA-1.5-7B and 13B, respectively, without HOPE-specific parameter tuning. Controlled ablations isolate the benefits of signed attribution and counterfactual verification beyond the number of edited examples alone.

1 INTRODUCTION

Multimodal large language models (MLLMs) achieve strong performance in visual question answering, image description, and multimodal dialogue (Liu et al., 2023; Li et al., 2023a), yet they still produce fluent responses that conflict with the visual input, including nonexistent objects, attributes, and relations (Rohrbach et al., 2018; Li et al., 2023b). Hallucination mitigation broadly relies on additional model alignment or training-free inference-time correction. The latter operates directly on pretrained MLLMs, typically by suppressing language priors, reinforcing visual signals, or modifying internal representations (Leng et al., 2024; Wang et al., 2025a; Zou et al., 2025; Lin et al., 2026). However, their interventions are often guided by proxy signals—such as attention, uncertainty, representation similarity, or predefined subspaces—***without explicitly verifying whether the selected visual evidence actually supports the current decision***.

This limitation exposes a central challenge: ***visual salience does not necessarily imply answer-specific decision relevance***. Gradient-based analyses can identify visual content that influences a model's prediction or uncertainty (Gröpl et al., 2026; Wang et al., 2026). Yet localization alone leaves two questions unresolved. First, an unsigned importance score does not indicate whether a visual token supports the current answer or a competing candidate. Second, even a candidate-conditioned attribution does not establish that perturbing the selected evidence will change the answer preference in the intended direction. This distinction is important since a model may attend to relevant visual evidence while still producing an incorrect answer (Liu et al., 2026). Without verifying how the current decision responds to the selected evidence, visual enhancement or representation editing may correct some hallucinations while corrupting well-grounded predictions,

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.

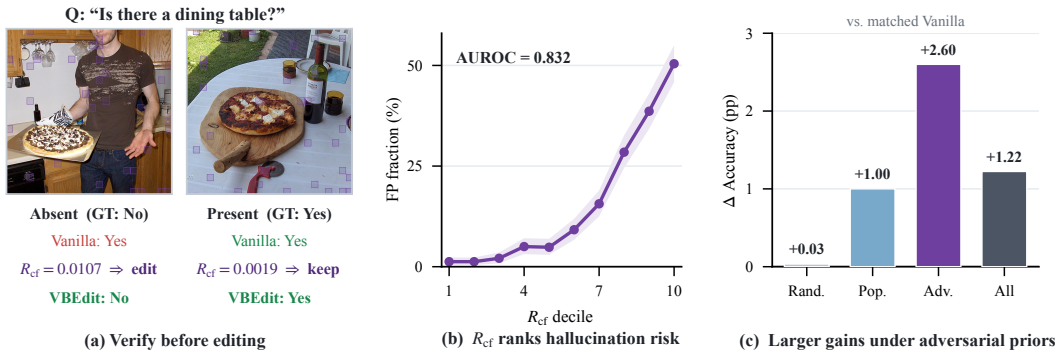

Figure 1: **Motivation for VBEDIT.** (a) Vanilla answers whether the queried object is absent or present; VBEDIT uses R_{cf} to edit only the hallucinated prediction. Purple boxes mark the source-grid positions of the top-32 visual tokens selected by signed attribution. (b) The false-positive fraction rises across R_{cf} deciles, showing that the counterfactual response ranks hallucination risk. (c) Accuracy gains over matched Vanilla across POPE splits.

leading to an unfavorable precision–recall trade-off. We therefore ask: ***can the decision relevance of visual evidence be verified before that evidence is used to guide an edit?***

Figure 1 illustrates this distinction. Panel (a) shows that the same question and positive response can be hallucinated in one visual context but grounded in another, calling for different editing decisions. In panel (b), the false-positive fraction rises sharply with R_{cf} (defined in Equation 10 of Section 3) among samples initially predicted as *Yes*, showing that the counterfactual response ranks hallucination risk. Panel (c) further shows that selective editing improves accuracy across all POPE splits, with the largest gain under adversarial language priors. Together, these observations motivate verifying the decision-level effect of localized evidence before editing the answer representation.

Accordingly, we introduce **VBEDIT**, a training-free "verify before edit" framework that tests answer-conditioned visual evidence before editing the first-step answer representation. VBEDIT represents the model's decision as a semantic margin between the current answer and competing candidates, and backpropagates this margin to obtain signed visual-token attributions that distinguish supporting from opposing evidence. It then counterfactually perturbs the most supportive tokens and measures the resulting margin change, yielding a sample-specific response R_{cf} that tests whether the localized evidence affects the decision. When this verification activates the gate, VBEDIT constructs a contrastive direction from supporting and opposing visual evidence and applies it only to the first-step answer representation.

Our contributions are summarized as follows:

- We formulate hallucination mitigation as a **verify-before-edit** problem and introduce **VBEDIT**, a training-free framework that conditionally edits the first-step answer representation after verifying the decision relevance of visual evidence.
- We develop answer-conditioned signed attribution to separate supporting from opposing visual evidence, together with a counterfactual response that tests the decision relevance of the selected tokens and enables reliability-gated contrastive editing.
- We conduct extensive evaluations across four MLLMs—LLaVA-1.5-7B/13B, Qwen3-VL-8B, and InternVL3.5-8B—and four benchmarks: POPE, MME, AMBER-D, and HOPE. VBEDIT consistently improves matched Vanilla on POPE and improves overall accuracy and F1 on AMBER-D at both LLaVA scales. Without HOPE-specific tuning, the corresponding POPE configurations raise HOPE accuracy by 5.93 and 9.43 points and Macro-F1 by 8.01 and 12.21 points on the 7B and 13B models, respectively.

2 RELATED WORK

2.1 MLLM HALLUCINATION MITIGATION

Multimodal large language models (MLLMs) often produce fluent responses that conflict with the visual input, particularly when language priors override weakly grounded visual evidence (Li et al.,

2023b). Mitigation methods broadly include training-based approaches, such as robust instruction tuning and hallucination-aware contrastive learning (Liu et al., 2024; Jiang et al., 2024), and inference-time approaches that leave the base model unchanged. Training-based methods can improve grounding at its source, but require curated data, additional supervision, and parameter updates. Inference-time methods instead modify decoding through visual contrast, attention penalties, intermediate-layer calibration, or gradient-estimated token influence (Leng et al., 2024; Huang et al., 2024; Wang et al., 2025a; 2026). They are plug-and-play, but contrastive variants may require extra forward passes, while broad logit correction can disturb otherwise correct decisions. Other methods amplify or reinject visual features (Yin et al., 2025; Zou et al., 2025; Zhu et al., 2026), edit intermediate representations or prefill caches (Suo et al., 2026; Zhang et al., 2026a; Lin et al., 2026), or localize hallucination-relevant layers and attention heads (Dang et al., 2026; Ji et al., 2026). Despite their effectiveness, these approaches commonly determine where or how to intervene from attention, uncertainty, learned directions, or aggregate representation statistics. Such signals can reveal correlation or sensitivity, but do not directly test whether the selected visual evidence is reliable for the current answer. VBEDIT instead follows a verify-before-edit principle: it edits only after a controlled counterfactual test establishes sample-specific decision relevance.

2.2 VISUAL ATTRIBUTION AND COUNTERFACTUAL EVIDENCE

Visual attribution aims to identify the image regions or visual tokens that influence a prediction. Recent methods backpropagate next-token uncertainty to visual embeddings (Gröpl et al., 2026), estimate token influence with first-order gradients (Wang et al., 2026), or adopt interventions to attribute hallucinations to particular layers (Dang et al., 2026). These estimators provide useful localization or sensitivity signals, but unsigned magnitude or dataset-level relevance does not indicate whether a feature supports or opposes a particular answer, nor whether perturbing it would change the current decision. Counterfactual reasoning offers a complementary way to evaluate evidence by measuring changes under controlled interventions. Prior work replaces image regions to explain predictions (Goyal et al., 2019), estimates modality-level effects (Zhou et al., 2025; Li et al., 2025), constructs diffusion-based counterfactual images to extract a global hallucination subspace (Dastmalchi et al., 2026), uses evidence interventions as a training signal (Zhang et al., 2026b), or removes visual tokens to form latent negative samples (Fa et al., 2026). In contrast, VBEDIT computes answer-conditioned signed attributions that separate supporting from opposing visual evidence, then measures how perturbing the selected tokens changes the current answer margin.

3 METHOD

3.1 OVERVIEW

Figure 2 illustrates the overall pipeline of VBEDIT. Given an image–query pair, a frozen MLLM produces the anchor-layer visual states $\mathcal{V}^{(\ell)} = \{\boldsymbol{v}_i^{(\ell)}\}_{i=1}^{N}$ and the first-step answer state $\boldsymbol{h}_1$. VBEDIT first attributes each visual token with respect to the semantic margin, separating evidence that supports the current answer from evidence that opposes it. It then counterfactually replaces the most supportive states and measures the resulting margin drop to test their decision relevance. If the verified evidence passes the reliability gate, VBEDIT edits $\boldsymbol{h}_1$ using a contrastive direction derived from the two evidence groups; otherwise, the state remains unchanged. Only the first-step answer state may be edited, while subsequent autoregressive decoding uses the original key–value cache.

3.2 ANSWER-CONDITIONED SEMANTIC MARGIN

For an image–question pair (I, q), let $\boldsymbol{z} \in \mathbb{R}^{|\mathcal{W}|}$ denote the token logits at the first answer position, where $\mathcal{W}$ is the output vocabulary. To account for surface-form variants of the same answer (e.g., `Yes` and `yes`), we associate each candidate $a \in \mathcal{A}$ with a token set $\mathcal{T}(a)$ and define

$$S(a) = \max_{t \in \mathcal{T}(a)} z_t, \qquad a^* = \arg\max_{a \in \mathcal{A}} S(a). \tag{1}$$

We quantify the confidence of the current answer relative to all competing candidates using the semantic margin

$$M(a^*) = S(a^*) - \log \sum_{b \in \mathcal{A} \setminus \{a^*\}} \exp\big(S(b)\big). \tag{2}$$

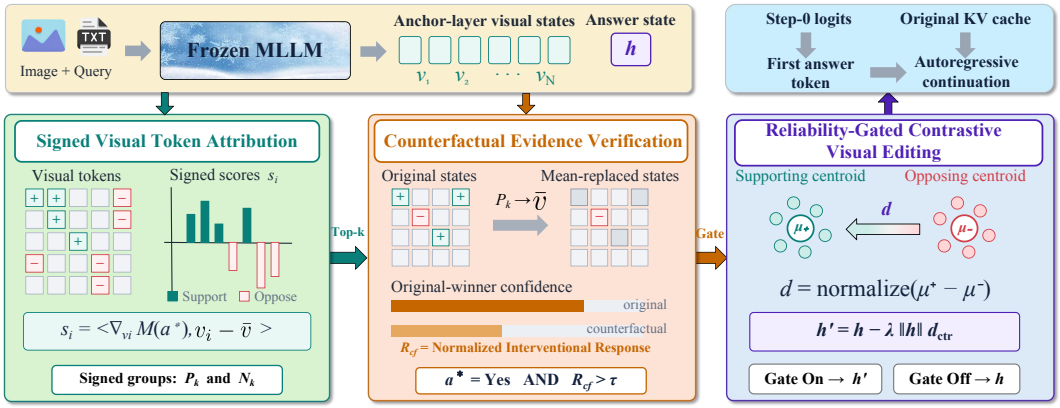

Figure 2: **Overview of VBEDIT.** The displayed condition $a^* = \text{Yes}$ is the binary object-hallucination instantiation; subsequent tokens are generated using the original key–value cache.

For a binary *Yes/No* task, this reduces to the winner–alternative semantic-score gap. Unlike entropy, which summarizes uncertainty over the output distribution, $M(a^*)$ measures confidence specifically in the current answer and provides the answer-conditioned objective for visual-token attribution.

3.3 SIGNED VISUAL-TOKEN ATTRIBUTION

Let $\boldsymbol{V}^{(\ell)} = \{\boldsymbol{v}_i^{(\ell)}\}_{i=1}^N$, with $\boldsymbol{v}_i^{(\ell)} \in \mathbb{R}^d$, denote the visual-token representations at anchor layer ℓ. We define their image-level mean and differentiate the semantic margin with respect to each token:

$$\bar{\boldsymbol{v}}^{(\ell)} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{v}_i^{(\ell)}, \qquad \boldsymbol{g}_i = \frac{\partial M(a^*)}{\partial \boldsymbol{v}_i^{(\ell)}}. \tag{3}$$

The signed attribution of token i is

$$s_i = \left\langle \boldsymbol{g}_i,\, \boldsymbol{v}_i^{(\ell)} - \bar{\boldsymbol{v}}^{(\ell)} \right\rangle. \tag{4}$$

Moving $\boldsymbol{v}_i^{(\ell)}$ toward the image-level mean has first-order margin change $-s_i$. Hence, $s_i > 0$ identifies evidence supporting the current answer, whereas $s_i < 0$ identifies opposing evidence. Unlike an unsigned gradient magnitude, s_i captures both sensitivity and direction with respect to the current semantic decision.

We retain up to k tokens with the strongest scores in each group:

$$\mathcal{P}_k = \mathrm{TopK}_k\big(s; \{i : s_i > 0\}\big), \qquad \mathcal{N}_k = \mathrm{TopK}_k\big(-s; \{i : s_i < 0\}\big), \tag{5}$$

where $\mathrm{TopK}_k(r; \mathcal{I})$ returns the indices of the largest ranking scores r_i over $\mathcal{I}$.

3.4 COUNTERFACTUAL EVIDENCE VERIFICATION

Signed attribution is a local, first-order quantity: even a large score does not establish that a finite perturbation of the selected evidence will materially affect the decision. VBEDIT therefore verifies the attributed support through an explicit counterfactual perturbation.

For each supportive token $i \in \mathcal{P}_k$, we construct

$$\tilde{\boldsymbol{v}}_i^{(\ell)} = (1-\alpha)\boldsymbol{v}_i^{(\ell)} + \alpha \bar{\boldsymbol{v}}^{(\ell)}, \qquad i \in \mathcal{P}_k, \tag{6}$$

while leaving all other visual tokens unchanged. Here $\alpha \in [0, 1]$ controls the replacement strength, with $\alpha = 1$ corresponding to full replacement by the image-level mean.

Proposition 1 (First-order optimality of signed visual-token selection) *Fix an answer candidate $a^\star$ and a decoder layer ℓ. Assume that the semantic margin $M(\boldsymbol{V}^{(\ell)}; a^\star)$ has an L-Lipschitz-continuous gradient in a neighborhood containing the relevant counterfactual perturbations. For*

each visual token, define

$$s_i = \left\langle \nabla_{\boldsymbol{v}_i^{(\ell)}} M(\boldsymbol{V}^{(\ell)}; a^\star), \boldsymbol{v}_i^{(\ell)} - \bar{\boldsymbol{v}}^{(\ell)} \right\rangle, \tag{7}$$

where $\bar{\boldsymbol{v}}^{(\ell)}$ is computed from the original visual-token representations and held fixed during the perturbation.

Given an intervention budget K and replacement strength $\alpha \in [0, 1]$, let $\mathcal{I}^\star$ contain up to K tokens with the largest positive values of s_i. Then $\mathcal{I}^\star$ maximizes the predicted first-order decrease of the semantic margin among all subsets satisfying $|\mathcal{I}| \le K$. Moreover, for any admissible subset $\mathcal{I}$,

$$\left| M(\widetilde{\boldsymbol{V}}_{\mathcal{I}}^{(\ell)}; a^\star) - M(\boldsymbol{V}^{(\ell)}; a^\star) + \alpha \sum_{i \in \mathcal{I}} s_i \right| \le \frac{L\alpha^2}{2} \sum_{i \in \mathcal{I}} \left\| \boldsymbol{v}_i^{(\ell)} - \bar{\boldsymbol{v}}^{(\ell)} \right\|_2^2. \tag{8}$$

Consequently, the true semantic margin is guaranteed to decrease whenever

$$\sum_{i \in \mathcal{I}} s_i > \frac{L\alpha}{2} \sum_{i \in \mathcal{I}} \left\| \boldsymbol{v}_i^{(\ell)} - \bar{\boldsymbol{v}}^{(\ell)} \right\|_2^2. \tag{9}$$

Proposition 1 justifies selecting the largest positive scores in Equation 5 and bounds the discrepancy between the predicted and realized margin changes. Because this guarantee is local, however, higher-order interactions may alter the response to a finite perturbation. We therefore measure the realized response while keeping the original winner a^* fixed. Let $M_{\text{orig}}(a^*)$ and $M_{\text{cf}}(a^*)$ denote its margins before and after the counterfactual perturbation. We define

$$R_{\text{cf}} = \text{clip}\left(\frac{M_{\text{orig}}(a^*) - M_{\text{cf}}(a^*)}{\max\left(|M_{\text{orig}}(a^*)|, \epsilon\right)}, 0, 1 \right), \tag{10}$$

where $\epsilon > 0$ ensures numerical stability near the decision boundary. A large R_{cf} indicates that mean-replacing the attributed support substantially reduces the current-answer margin. Thus, R_{cf} measures sample-specific decision dependence rather than serving as a calibrated correctness probability or a formal causal effect.

3.5 Reliability-gated contrastive visual editing

The signed attribution scores yield supportive and opposing evidence sets, which we summarize by their centroids:

$$\boldsymbol{\mu}^+ = \frac{1}{|\mathcal{P}_k|} \sum_{i \in \mathcal{P}_k} \boldsymbol{v}_i^{(\ell)}, \qquad \boldsymbol{\mu}^- = \frac{1}{|\mathcal{N}_k|} \sum_{i \in \mathcal{N}_k} \boldsymbol{v}_i^{(\ell)}. \tag{11}$$

Their normalized difference defines the contrastive visual direction

$$\boldsymbol{u}_{\text{ctr}} = \frac{\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-}{\|\boldsymbol{\mu}^+ - \boldsymbol{\mu}^-\|_2 + \epsilon}. \tag{12}$$

Unlike a support-only representation, this direction captures the separation between evidence supporting and opposing the current answer. Because the sign of a representation-space direction is not intrinsic, we orient $\boldsymbol{u}_{\text{ctr}}$ by its first-order effect on the semantic margin. Let $\boldsymbol{h}_1$ denote the hidden state used to predict the first answer token. We define

$$\rho = \text{sign}\left(\left\langle \frac{\partial M(a^*)}{\partial \boldsymbol{h}_1}, \boldsymbol{u}_{\text{ctr}} \right\rangle \right), \qquad \boldsymbol{d}_{\text{ctr}} = \rho\, \boldsymbol{u}_{\text{ctr}}, \tag{13}$$

so that moving along $\boldsymbol{d}_{\text{ctr}}$ locally increases the margin of the current answer.

VBEdit applies the edit only when the current answer belongs to the targeted risk set and its counterfactual response exceeds threshold τ:

$$G = \mathbb{I}\left[a^* \in \mathcal{A}_{\text{risk}}\right] \cdot \mathbb{I}\left[R_{\text{cf}} > \tau\right], \tag{14}$$

Here, $\mathcal{A}_{\text{risk}}$ specifies the candidates eligible for intervention. For binary visual queries, we use $\mathcal{A}_{\text{risk}} = \{\textit{Yes}\}$, while the formulation can accommodate broader candidate sets.

The edited first-step representation is

$$\boldsymbol{h}_1' = \boldsymbol{h}_1 - G\lambda \|\boldsymbol{h}_1\|_2\, \boldsymbol{d}_{\text{ctr}}, \tag{15}$$

where $\lambda \ge 0$ is the relative editing strength. Scaling by $\|\boldsymbol{h}_1\|_2$ adapts the edit magnitude to the hidden-state scale across samples and model sizes. Consequently, $G = 0$ exactly recovers the original model, whereas $G = 1$ moves the state against the visual direction supporting the targeted answer and thereby reduces its semantic margin.

Table 1: Object hallucination evaluation on the full POPE benchmark. Baseline values are reported results from prior works. Best and second-best values are bolded and underlined, respectively.

Category	Method	LLaVA-1.5-7B		LLaVA-1.5-13B		Qwen3-VL-8B	
		Accuracy↑	F1↑	Accuracy↑	F1↑	Accuracy↑	F1↑
Random	VCD	88.40	87.70	88.90	87.80	90.67	90.83
	ICD	88.10	87.60	88.10	87.60	91.74	91.75
	VAF	89.60	89.30	90.10	89.90	–	–
	DeCo	89.10	89.80	82.20	84.90	<u>91.96</u>	<u>92.36</u>
	HulluEdit	<u>90.40</u>	**90.50**	<u>90.60</u>	<u>90.80</u>	–	–
	VBEDIT	**90.58**	<u>90.43</u>	**91.10**	**91.02**	**92.92**	**92.82**
Popular	VCD	83.10	84.10	83.70	85.10	83.27	84.53
	ICD	82.10	82.90	82.90	84.30	84.80	<u>85.90</u>
	VAF	84.50	84.90	85.20	86.40	–	–
	DeCo	84.60	85.80	78.40	81.70	<u>86.01</u>	83.81
	HulluEdit	<u>87.50</u>	<u>87.60</u>	<u>88.00</u>	<u>88.30</u>	–	–
	VBEDIT	**88.20**	**87.97**	**89.40**	**89.20**	**89.10**	**89.07**
Adversarial	VCD	78.10	79.60	78.20	79.70	81.36	83.60
	ICD	78.50	79.90	79.10	80.10	82.52	<u>83.89</u>
	VAF	80.10	81.00	80.70	81.70	–	–
	DeCo	78.30	81.10	72.60	77.90	<u>82.74</u>	83.81
	HulluEdit	<u>82.50</u>	<u>83.40</u>	<u>82.70</u>	<u>84.00</u>	–	–
	VBEDIT	**84.23**	**84.55**	**85.57**	**85.85**	**86.90**	**87.15**

4 EXPERIMENTS

4.1 EXPERIMENTAL SETUPS

We evaluate object hallucination on the full POPE benchmark (Li et al., 2023b), which comprises the Random, Popular, and Adversarial splits, and report Accuracy and F1. We also report the four object-centric MME perception tasks—Existence, Count, Position, and Color—using their task scores (Fu et al., 2025). We additionally evaluate the discriminative subset of AMBER (abbreviated as AMBER-D) (Wang et al., 2023), comprising 14,216 Yes/No questions about object existence, attributes, and relations. To examine cross-benchmark transfer, we use the public description-based set of HOPE (Xie et al., 2025). The main POPE and MME evaluation covers LLaVA-1.5-7B, LLaVA-1.5-13B (Liu et al., 2023), and Qwen3-VL-8B (Bai et al., 2025). Additional cross-model and cross-benchmark results, including POPE on InternVL3.5-8B (Wang et al., 2025b), AMBER-D on both LLaVA scales, and HOPE on both LLaVA scales, are reported in Appendix A.3. Note that the editing affects only the first answer step and subsequent generation uses the original key–value cache. The main POPE and MME scores are based on complete greedy generated responses. Further implementation and evaluation details are provided in Appendix A.2.

4.2 MAIN RESULTS ON POPE AND CROSS-BENCHMARK TRANSFER TO MME

Tables 1 and 2 compare VBEDIT with representative hallucination-mitigation baselines on POPE and MME. Method descriptions are provided in Appendix A.2.3. On LLaVA-1.5-7B, VBEDIT attains the highest accuracy on all splits and improves both metrics over HulluEdit on Popular and Adversarial; only Random F1 is 0.07 points lower. On LLaVA-1.5-13B, it surpasses the reported HulluEdit across all six metrics, including gains of 2.87 and 1.85 points in Adversarial Accuracy and F1. HulluEdit does not report Qwen3-VL-8B, where VBEDIT leads all reported baselines listed in the table on every split. The largest gains consistently occur on Adversarial, supporting verified visual editing under stronger language priors. For MME, we apply the corresponding POPE-selected configurations without MME-specific parameter search. On LLaVA-1.5-7B, VBEDIT matches the published HulluEdit result on Existence and improves Count, Position, and Color by 25.00, 18.33, and 10.00 points. Relative to matched Vanilla, it raises the total score from 573.33 to 655.00 on LLaVA-1.5-13B and from 710.00 to 723.33 on Qwen3-VL-8B. Overall, the gains are concentrated in counting, spatial reasoning, and attribute recognition.

Table 2: Fine-grained MME results. VBEDIT employs configurations selected on POPE without MME-specific parameter tuning.

Method	Existence↑	Count↑	Position↑	Color↑	Total Score
LLaVA-1.5-7B	181.67	118.33	104.44	152.78	557.22
Nullu	190.00	121.11	105.56	156.67	573.34
DeCo	175.00	128.33	98.33	125.00	526.66
HulluEdit	**195.00**	105.00	126.67	160.00	586.67
VBEDIT	**195.00**	**130.00**	**145.00**	**170.00**	**640.00**
LLaVA-1.5-13B	190.00	140.00	103.33	140.00	573.33
VBEDIT	**195.00**	**151.67**	**138.33**	**170.00**	**655.00**
Qwen3-VL-8B	**195.00**	176.67	148.33	190.00	710.00
VBEDIT	**195.00**	**180.00**	**153.33**	**195.00**	**723.33**

Table 3: Component ablations on full POPE with LLaVA-1.5-7B. ΔCorrect measures beneficial FP→TN transitions minus harmful TP→FN transitions.

Variant / control	Acc.↑	Prec.↑	F1↑	FP→TN	TP→FN	ΔCorrect↑
Vanilla	86.42	84.41	86.96	0	0	0
Random visual-token attribution	86.42	84.42	86.96	0.6	0.8	-0.2
w/o signed attribution	86.42	84.50	86.94	6	6	0
w/o counterfactual centering	86.45	84.69	86.95	18	15	+3
Matched-random reliability gate	86.67	85.67	87.01	79.8	57.4	+22.4
w/o contrastive direction	87.08	**89.26**	86.87	287	228	+59
Accuracy-selected Yes-bias control	87.38	87.85	87.46	203	117	+86
Edit all eligible (w/o verification)	87.54	89.24	87.41	280	180	+100
VBEDIT	**87.64**	88.93	**87.58**	262	153	**+109**

Computational overhead and efficiency. On a fixed POPE subset, VBEDIT has $1.81\times$ the mean latency of matched Vanilla, with 36.7 MiB additional peak allocated memory. Most of the latency overhead comes from signed attribution and counterfactual verification (Appendix A.5).

4.3 ABLATION STUDIES

We conduct several ablation studies on the full POPE benchmark using LLaVA-1.5-7B. Unless stated otherwise, all variants share the same model, prompt, and generation protocol; only the component under study is changed. Randomized controls average five predetermined seeds.

4.3.1 CONTRIBUTION OF INDIVIDUAL COMPONENTS

We ablate each component while retaining the rest of the pipeline. *Random visual-token attribution* selects equally many random tokens, whereas *unsigned attribution* removes the sign of the margin gradient. The remaining variants omit centering, reliability gating, or the opposing-evidence prototype. The *matched-random gate* preserves VBEDIT's per-split editing budget while randomizing the edited samples. Table 3 shows that VBEDIT improves accuracy from 86.42% to 87.64% and F1 from 86.96% to 87.58%. Random-token and unsigned variants remain near Vanilla, while removing centering reduces ΔCorrect to $+3$. Under the same editing budget, matched-random gating achieves only $+22.4$, compared with $+109$ for VBEDIT, confirming that counterfactual verification identifies more useful interventions. The accuracy-selected Yes-bias recovers part of the gain ($\Delta\text{Correct} = +86$), but remains below VBEDIT in both Accuracy and F1, suggesting that uniform Yes suppression does not fully reproduce the gain. Editing all eligible predictions reaches $+100$ but introduces more harmful suppressions, and removing the opposing-evidence prototype further increases such errors. These results demonstrate the complementary roles of signed attribution, centering, reliability gating, and contrastive editing. Full results are provided in Appendix A.4.

Table 4: Attribution and counterfactual evidence quality on full POPE.

Evidence estimator	First-order agreement (%)↑	$\mathrm{AUC}_{\mathrm{FP}>\mathrm{TP}}$ (%)↑	FP→TN	TP→FN	ΔCorrect↑
Random visual tokens	78.88	37.18	0.6	0.8	-0.2
Entropy-gradient magnitude	77.36	38.96	6	0	+6
Unsigned margin-gradient	78.96	42.60	6	6	0
Signed-margin attribution	**100.00**	**83.16**	262	153	**+109**

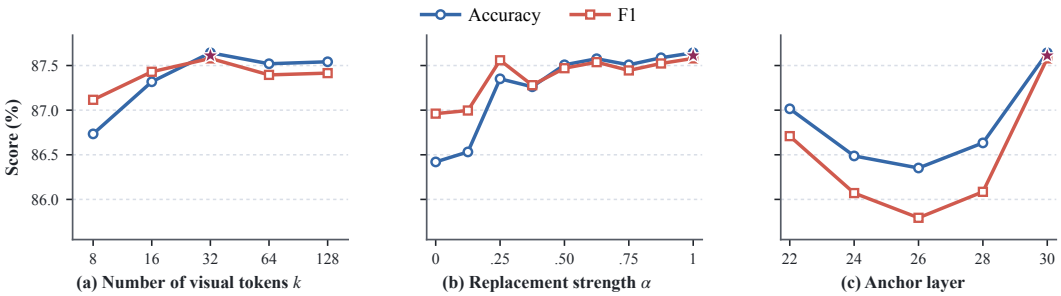

Figure 3: Generated-text sensitivity of VBEDIT on full POPE with LLaVA-1.5-7B. We vary (a) the number of attributed visual tokens, (b) the counterfactual replacement strength, and (c) the anchor layer, while fixing all remaining parameters. The selected operating point is marked with a star.

4.3.2 ATTRIBUTION AND COUNTERFACTUAL EVIDENCE QUALITY

We evaluate whether localized evidence is decision-relevant using *first-order agreement*, which compares the sign of the observed answer-margin change with that predicted by attribution. We also test whether the counterfactual response ranks false positives above true positives among Vanilla-predicted positive samples, denoted by $\mathrm{AUC}_{\mathrm{FP}>\mathrm{TP}}$. Ground truth is used only for this post-hoc evaluation. Table 4 shows that signed-margin attribution is the only estimator combining perfect directional agreement, strong FP–TP discriminability, and a substantial positive ΔCorrect. Its 83.16% AUC contrasts with the below-chance alternatives. The 100% agreement indicates local directional faithfulness, not perfect semantic grounding; likewise, a large response indicates strong dependence on concentrated, potentially fragile visual features rather than factual correctness. VBEDIT therefore treats it as an intervention signal, not a correctness label.

4.3.3 SENSITIVITY TO INTERVENTION CONFIGURATIONS

We vary three structural hyperparameters—the number of selected visual tokens k, counterfactual replacement strength α, and anchor layer with all remaining settings fixed. Figure 3(a) identifies $k = 32$ as the best setting. Larger k improves FP–TP AUC and corrects more false positives but exposes more true positives to harmful suppression, whereas smaller k provides insufficient evidence coverage. Panel (b) shows stable gains under moderate-to-strong replacement, with $\alpha = 1$ performing best; exact Vanilla recovery at $\alpha = 0$ verifies the intervention boundary. Panel (c) shows that earlier layers cause substantially more harmful suppression, indicating that localization alone is insufficient without an effective answer-level editing site.

4.4 UNDERSTANDING VBEDIT'S VERIFICATION AND EDITING BEHAVIOR

We next illustrate the resulting verify-before-edit decisions at the sample level, and examine how reliability gating and editing strength translate this diagnostic signal into prediction corrections.

Sample-level verify-before-edit behavior. Figure 4 complements the aggregate analysis with two matched-query examples. Although signed attribution identifies supporting visual tokens in both cases, only the hallucinated answer exhibits a counterfactual response sufficient to activate the gate. VBEDIT therefore reverses the false-positive margin while leaving the grounded prediction unchanged, illustrating selective verification before editing.

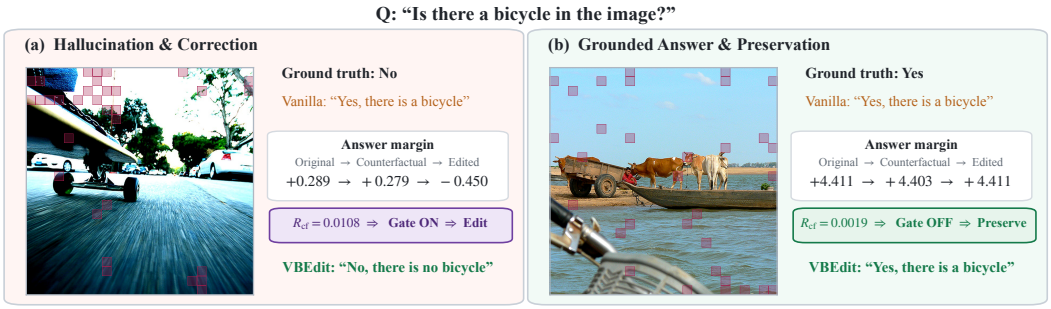

Figure 4: Sample-level verify-before-edit behavior for two examples sharing the same question and Vanilla answer. A large counterfactual response activates editing and corrects the hallucination (a), whereas a small response preserves the grounded prediction (b). Red cells mark the source-grid positions of the top-32 contextualized visual tokens selected by signed attribution. Margins correspond to the original, counterfactual-probe, and post-gate decisions; gate-off leaves the original margin unchanged. Editing acts on hidden states, not image pixels.

Selective gating improves correction efficiency. Figure 5(a) varies the gate threshold at the selected editing strength $\lambda = 1.1$. Editing all 4,788 samples predicted as *Yes* by the candidate readout increases the number of correct predictions by 102, whereas the selected gate edits only 1,444 samples (30.2%) and increases it by 111. For this predicted-Yes intervention, the correctness change equals FP→TN − TP→FN. The gate therefore removes many low-utility interventions while retaining the samples on which editing is most beneficial.

Editing strength controls the precision–recall trade-off. Figure 5(b) shows that increasing λ generally raises precision but lowers recall. Accuracy consequently follows a non-monotonic trend, with $\lambda = 1.1$ among the best-performing sampled strengths. At this selected configuration, complete generated-text evaluation confirms a positive correction balance: VBEDIT corrects 262 false positives while converting 153 true positives into false negatives, increasing the number of correct predictions by 109. Selective gating determines *where* to intervene, while the editing strength controls *how strongly* the answer representation is modified.

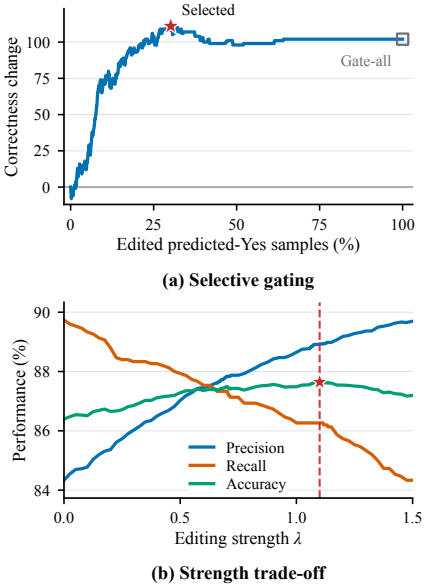

Figure 5: VBEDIT intervention behavior on POPE: (a) correctness change versus gate coverage at $\lambda = 1.1$; (b) precision–recall trade-off.

5 CONCLUSION

We introduced **VBEDIT**, a training-free framework following the ***verify-before-edit*** principle for mitigating hallucinations in MLLMs. VBEDIT uses signed visual-token attribution to identify supporting and opposing evidence, verifies its decision relevance through counterfactual perturbation, and selectively edits the answer representation along a contrastive visual direction. Experiments across multiple MLLM architectures demonstrate consistent improvements over the matched Vanilla and representative methods. The same frozen configurations also improve MME, AMBER-D, and HOPE without dataset-specific tuning, supporting cross-benchmark transferability. Ablations further confirm the contributions of signed attribution, counterfactual verification, and selective editing.

Limitations. First, VBEDIT depends on visual evidence already encoded by the frozen MLLM and cannot recover missing representations. Second, although VBEDIT applies naturally to binary visual queries, its extension to multi-candidate and open-ended generation remains unexplored. Third, R_{cf} measures sensitivity to internal visual-token perturbations, not image-level causality or pixel-level localization.

AI USE STATEMENT

In this work, we used generative AI tools to assist manuscript preparation, including editing the manuscript to improve grammar, readability, clarity, and stylistic consistency. All AI-assisted suggestions and edits were reviewed and verified by the authors. We take responsibility for the final content of this work, including all text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies hallucination mitigation in pretrained multimodal large language models using publicly available models and benchmarks. It does not involve human subjects or collect private or sensitive user data. Although reducing visually unsupported answers may improve model reliability, VBEDIT does not provide a guarantee of factuality or safety and may retain errors and biases inherited from the underlying models and data. Our claims are therefore limited to the evaluated models and tasks; deployment in high-stakes settings would require independent validation and appropriate safeguards. This work adheres to the ICLR Code of Ethics.

REPRODUCIBILITY STATEMENT

Section 3 specifies the semantic margin, signed attribution, counterfactual verification, reliability gate, and first-step editing procedure. Section 4 describes the evaluated models, benchmarks, metrics, and evaluation protocols. Appendix A.2 provides checkpoint identifiers, input and generation settings, evaluation procedures, and the intervention configurations used for each model. Appendix A.1 provides the complete proof of Proposition 1, and Appendix A.4 reports additional controls and complete sensitivity results. Upon acceptance, we plan to release the implementation, configuration files, evaluation scripts, and instructions needed to reproduce the reported experiments.

REFERENCES

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*, 2025. URL https://arxiv.org/abs/2511.21631.

Xinlong Chen, Yuanxing Zhang, Qiang Liu, Junfei Wu, Fuzheng Zhang, and Tieniu Tan. Mixture of decoding: An attention-inspired adaptive decoding strategy to mitigate hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 8525–8542, 2025.

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.

Hamidreza Dastmalchi, Aijun An, Ali Cheraghian, and Hamed Barzamini. Fighting hallucinations with counterfactuals: Diffusion-guided perturbations for LVLM hallucination suppression. *arXiv preprint arXiv:2603.10470*, 2026. URL https://arxiv.org/abs/2603.10470.

Zhan Fa, Yue Duan, Jian Zhang, Lei Qi, and Yinghuan Shi. One token, two fates: A unified framework via vision token manipulation against MLLMs hallucination. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11106–11115, 2026.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination con-

trol by visual information grounding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14303–14312, 2024.

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.

Yash Goyal, Ziyan Wu, Jan Ernst, Dhruv Batra, Devi Parikh, and Stefan Lee. Counterfactual visual explanations. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 2376–2384. PMLR, 2019.

Marcel Gröpl, Jaewoo Jung, Seungryong Kim, Marc Pollefeys, and Sunghwan Hong. Entropy-gradient grounding: Training-free evidence retrieval in vision-language models. *arXiv preprint arXiv:2604.08456*, 2026. URL https://arxiv.org/abs/2604.08456.

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

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

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.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27036–27046, 2024.

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.

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, pp. 19730–19742. PMLR, 2023a.

Li Li, Jiashu Qu, Linxin Song, Yuxiao Zhou, Yuehan Qin, Tiankai Yang, and Yue Zhao. Treble counterfactual VLMs: A causal approach to hallucination. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 18423–18434. Association for Computational Linguistics, 2025.

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, 2023b.

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*, pp. 57689–57733, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 34892–34916, 2023.

Zhining Liu, Ziyi Chen, Hui Liu, Chen Luo, Xianfeng Tang, Suhang Wang, Jingying Zeng, Zhenwei Dai, Zhan Shi, Tianxin Wei, et al. Seeing but not believing: Probing the disconnect between visual attention and answer correctness in vlms. In *International Conference on Learning Representations*, volume 2026, pp. 72849–72876, 2026.

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.

Wei Suo, Hanzu Zhang, Lijun Zhang, Ji Ma, Peng Wang, and Yanning Zhang. Hallucination-aware intermediate representation edit in large vision-language models. In *The Fourteenth International Conference on Learning Representations*, 2026.

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*, pp. 13712–13736, 2025a.

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

Shan Wang, Maying Shen, Nadine Chang, Chuong Nguyen, Hongdong Li, and Jose M. Alvarez. Mitigating multimodal hallucinations via gradient-based self-reflection. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 32539–32549, 2026.

Weiyun Wang, Zhangwei Gao, Lixin Gu, et al. InternVL3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025b. doi: 10.48550/arXiv.2508.18265. URL https://arxiv.org/abs/2508.18265.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the association for computational linguistics: ACL 2024*, pp. 15840–15853, 2024.

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

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

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. Nullu: Mitigating object hallucinations in large vision-language models via HalluSpace projection. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14635–14645, 2025.

Hao Yin, Guangzong Si, and Zilei Wang. ClearSight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14625–14634, 2025.

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, 2026a.

Yan Zhang, Zhijin Qin, Feng Xu, Guiguang Ding, and Jungong Han. CofactGVR: Counterfactual intervention for grounded visual reasoning. In *Forty-third International Conference on Machine Learning*, 2026b. URL https://openreview.net/forum?id=anm4mC6KLx.

Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. In *International Conference on Learning Representations*, pp. 54415–54439, 2025.

Xingyu Zhu, Kesen Zhao, Liang Yi, Shuo Wang, Zhicai Wang, Beier Zhu, Hanwang Zhang, and Xiangnan He. Look carefully: Adaptive visual reinforcements in multimodal large language models for hallucination mitigation. In *International Conference on Learning Representations*, pp. 81235–81256, 2026.

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and Xuming Hu. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 80873–80899. PMLR, 2025.

A TECHNICAL APPENDICES AND SUPPLEMENTARY MATERIAL

This appendix provides theoretical derivations, reproducibility details, cross-architecture results, additional diagnostics, qualitative analyses, efficiency measurements, and an extended discussion of limitations. It is organized as follows:

- Section A.1 provides the **complete proof of Proposition 1**.
- Section A.2 documents **implementation and evaluation details**.
- Section A.3 reports additional results across models and benchmarks, including **POPE results on InternVL3.5-8B** (Section A.3.1), **AMBER results on LLaVA-1.5-7B/13B** (Section A.3.2), and **HOPE results** (Section A.3.3).
- Section A.4 presents **additional ablations and complete sensitivity results**.
- Section A.5 reports **computational overhead and efficiency**.
- Section A.6 presents **qualitative analyses** of visual evidence, successful corrections, preserved predictions, and representative failure cases.
- Section A.7 provides **limitations and future directions**.

A.1 PROOF OF PROPOSITION 1

For a selected set of visual tokens $\mathcal{I}$, VBEDIT constructs the counterfactual representations as

$$\widetilde{\boldsymbol{v}}_i^{(\ell)} = \begin{cases} (1-\alpha)\boldsymbol{v}_i^{(\ell)} + \alpha\bar{\boldsymbol{v}}^{(\ell)}, & i \in \mathcal{I}, \\ \boldsymbol{v}_i^{(\ell)}, & i \notin \mathcal{I}. \end{cases} \tag{16}$$

The corresponding perturbation is therefore

$$\Delta\boldsymbol{v}_i^{(\ell)} = \widetilde{\boldsymbol{v}}_i^{(\ell)} - \boldsymbol{v}_i^{(\ell)} = -\alpha\left(\boldsymbol{v}_i^{(\ell)} - \bar{\boldsymbol{v}}^{(\ell)}\right), \qquad i \in \mathcal{I}, \tag{17}$$

and $\Delta\boldsymbol{v}_i^{(\ell)} = \mathbf{0}$ for $i \notin \mathcal{I}$. Let $\Delta\boldsymbol{V}_{\mathcal{I}}^{(\ell)}$ denote the matrix collecting these token-wise perturbations. It follows that

$$\left\|\Delta\boldsymbol{V}_{\mathcal{I}}^{(\ell)}\right\|_F^2 = \alpha^2 \sum_{i\in\mathcal{I}} \left\|\boldsymbol{v}_i^{(\ell)} - \bar{\boldsymbol{v}}^{(\ell)}\right\|_2^2. \tag{18}$$

For compactness, define the scalar function

$$\phi(t) = M\left(\boldsymbol{V}^{(\ell)} + t\Delta\boldsymbol{V}_{\mathcal{I}}^{(\ell)}; a^\star\right), \qquad t \in [0,1]. \tag{19}$$

The answer candidate $a^\star$ is held fixed along this path. By the fundamental theorem of calculus,

$$\begin{aligned} \phi(1) - \phi(0) &= \int_0^1 \phi'(t)\,\mathrm{d}t \\ &= \int_0^1 \left\langle \nabla_{\boldsymbol{V}^{(\ell)}} M\left(\boldsymbol{V}^{(\ell)} + t\Delta\boldsymbol{V}_{\mathcal{I}}^{(\ell)}; a^\star\right), \Delta\boldsymbol{V}_{\mathcal{I}}^{(\ell)} \right\rangle_F \mathrm{d}t. \end{aligned} \tag{20}$$

Adding and subtracting the gradient at the unperturbed representation gives

$$\phi(1) - \phi(0) = \left\langle \nabla_{\boldsymbol{V}^{(\ell)}} M(\boldsymbol{V}^{(\ell)}; a^\star), \Delta\boldsymbol{V}_{\mathcal{I}}^{(\ell)} \right\rangle_F + \varepsilon_{\mathcal{I}}, \tag{21}$$

where

$$\begin{aligned} \varepsilon_{\mathcal{I}} = \int_0^1 \Big\langle & \nabla_{\boldsymbol{V}^{(\ell)}} M\left(\boldsymbol{V}^{(\ell)} + t\Delta\boldsymbol{V}_{\mathcal{I}}^{(\ell)}; a^\star\right) \\ & - \nabla_{\boldsymbol{V}^{(\ell)}} M(\boldsymbol{V}^{(\ell)}; a^\star), \Delta\boldsymbol{V}_{\mathcal{I}}^{(\ell)} \Big\rangle_F \mathrm{d}t. \end{aligned} \tag{22}$$

Because the gradient of M is L-Lipschitz continuous, the Cauchy–Schwarz inequality gives

$$\begin{aligned} |\varepsilon_{\mathcal{I}}| &\le \int_0^1 Lt \left\| \Delta \boldsymbol{V}_{\mathcal{I}}^{(\ell)} \right\|_F^2 \mathrm{d}t \\ &= \frac{L}{2} \left\| \Delta \boldsymbol{V}_{\mathcal{I}}^{(\ell)} \right\|_F^2 \\ &= \frac{L\alpha^2}{2} \sum_{i \in \mathcal{I}} \left\| \boldsymbol{v}_i^{(\ell)} - \bar{\boldsymbol{v}}^{(\ell)} \right\|_2^2 . \end{aligned} \tag{23}$$

The linear term in Equation 21 decomposes over visual tokens:

$$\begin{aligned} &\left\langle \nabla_{\boldsymbol{V}^{(\ell)}} M(\boldsymbol{V}^{(\ell)}; a^\star), \Delta \boldsymbol{V}_{\mathcal{I}}^{(\ell)} \right\rangle_F \\ &= \sum_{i \in \mathcal{I}} \left\langle \nabla_{\boldsymbol{v}_i^{(\ell)}} M(\boldsymbol{V}^{(\ell)}; a^\star), \Delta \boldsymbol{v}_i^{(\ell)} \right\rangle \\ &= -\alpha \sum_{i \in \mathcal{I}} \left\langle \nabla_{\boldsymbol{v}_i^{(\ell)}} M(\boldsymbol{V}^{(\ell)}; a^\star), \boldsymbol{v}_i^{(\ell)} - \bar{\boldsymbol{v}}^{(\ell)} \right\rangle \\ &= -\alpha \sum_{i \in \mathcal{I}} s_i . \end{aligned} \tag{24}$$

Combining Equations 21, 23, and 24 yields

$$M(\widetilde{\boldsymbol{V}}_{\mathcal{I}}^{(\ell)}; a^\star) = M(\boldsymbol{V}^{(\ell)}; a^\star) - \alpha \sum_{i \in \mathcal{I}} s_i + \varepsilon_{\mathcal{I}}, \tag{25}$$

with the remainder bounded as stated in Equation 8.

It remains to establish the optimality of positive top-K selection. Since $\alpha \ge 0$, minimizing the linearized counterfactual margin is equivalent to maximizing

$$\sum_{i \in \mathcal{I}} s_i \quad \text{subject to} \quad |\mathcal{I}| \le K. \tag{26}$$

Any token with $s_i \le 0$ cannot improve this objective: removing such a token leaves the sum unchanged or increases it. Hence an optimal set contains only positively scored tokens. Among positive scores, suppose an admissible set contains token j but excludes token i with $s_i > s_j$. Replacing j with i strictly increases the objective. Repeating this exchange yields the set containing the largest $\min(K, |\{i : s_i > 0\}|)$ positive scores. This is precisely the positive top-K selection used by VBEDIT.

Finally, Equation 25 and the upper bound on $\varepsilon_{\mathcal{I}}$ imply

$$\begin{aligned} &M(\widetilde{\boldsymbol{V}}_{\mathcal{I}}^{(\ell)}; a^\star) - M(\boldsymbol{V}^{(\ell)}; a^\star) \\ &\le -\alpha \sum_{i \in \mathcal{I}} s_i + \frac{L\alpha^2}{2} \sum_{i \in \mathcal{I}} \left\| \boldsymbol{v}_i^{(\ell)} - \bar{\boldsymbol{v}}^{(\ell)} \right\|_2^2 . \end{aligned} \tag{27}$$

Therefore, the semantic margin decreases strictly whenever Equation 9 holds.

Implication for counterfactual verification. The proposition establishes a local, first-order optimality result; it does not assert that the semantic margin must decrease for an arbitrarily large intervention. Curvature, nonlinear attention interactions, and downstream normalization may cause the higher-order remainder to dominate the linear term. VBEDIT therefore measures the actual margin response after counterfactual replacement. This verification step tests whether the first-order attribution remains predictive under the finite perturbation used at inference time, preventing attribution magnitude alone from being treated as reliable evidence.

A.2 Implementation and Evaluation Details

A.2.1 Models and Benchmarks

We employ the public instruction-tuned MLLM checkpoints `liuhaotian/llava-v1.5-7b`, `liuhaotian/llava-v1.5-13b`, `Qwen/Qwen3-VL-8B-Instruct`, and `OpenGVLab/`

`InternVL3_5-8B-Instruct`. Each checkpoint is evaluated with its official tokenizer, image processor, and conversation template.

For POPE, we use the complete 8,910-example evaluation set and compute Accuracy, Precision, Recall, and F1 from the generated answers; the main comparison table reports Accuracy and F1 for compactness (Table 1). For MME, we follow the official paired-question evaluation and report the four object-centric perception tasks—Existence, Count, Position, and Color—in the main paper. For AMBER (Wang et al., 2023), we evaluate the complete discriminative subset of 14,216 Yes/No questions covering existence, attribute, and relation hallucinations, and report Accuracy and Overall F1 using the official evaluation script. HOPE contains 3,000 balanced Yes/No questions over 500 MS-COCO images. Notably, the configurations of VBEDIT, HulluEdit (Lin et al., 2026), and VCD (Leng et al., 2024) are fixed on POPE and applied without HOPE-specific parameter selection under a common greedy generated-text protocol.

A.2.2 INFERENCE AND EDITING PROTOCOL

We use greedy decoding and retain the complete generated response rather than evaluating candidate logits alone. All model parameters remain frozen. Gradients are taken only with respect to anchor-layer activations, and parameter gradients are neither required nor accumulated. Visual-token selection is restricted to the detected visual span, with the effective top-k capped by the number of available visual tokens. Counterfactual replacement and hidden-state editing affect only the first answer step. Subsequent autoregressive decoding reuses the original prompt-side key–value cache and does not invoke the vision encoder again. When the reliability gate is inactive, the first token, continuation, and parsed answer are required to match Vanilla exactly. Unless the varied factor is stated explicitly, the LLaVA-1.5-7B POPE analyses and ablations use $(\ell, k, \alpha, \lambda) = (30, 32, 1.0, 1.1)$. The ablation and sensitivity experiments retain this configuration while changing only the component or hyperparameter under study. Randomized controls use predetermined seeds and never select a seed according to test performance. All experiments are conducted on a single NVIDIA GeForce RTX 5090 GPU.

A.2.3 COMPARED METHODS

We compare VBEDIT with greedy decoding and representative post-hoc hallucination-mitigation methods. Visual Contrastive Decoding (VCD) contrasts predictions conditioned on the original and distorted images (Leng et al., 2024), whereas Instruction Contrastive Decoding (ICD) contrasts standard and hallucination-inducing instructions (Wang et al., 2024). Visual Amplification Fusion (VAF) strengthens mid-layer visual representations (Yin et al., 2025). DeCo dynamically calibrates decoding using intermediate-layer evidence (Wang et al., 2025a), while HulluEdit performs single-pass evidence-consistent subspace editing (Lin et al., 2026). For MME, we additionally report Nullu, which projects model weights away from a learned hallucination subspace (Yang et al., 2025).

A.3 ADDITIONAL RESULTS ACROSS MODELS AND BENCHMARKS

A.3.1 ADDITIONAL POPE RESULTS ON INTERNVL3.5-8B

We additionally evaluate InternVL3.5-8B-Instruct on the 8,910-example POPE benchmark. Table 5 compares VBEDIT with its matched Vanilla baseline and with the FINER-Tuning results (Xiao et al., 2026). Matched Vanilla and VBEDIT use the same checkpoint, prompt, and greedy generated-text protocol. Relative to matched Vanilla, VBEDIT improves Accuracy by 0.17, 0.37, and 0.30 points and F1 by 0.18, 0.36, and 0.30 points on Random, Popular, and Adversarial, respectively. The matched results support the cross-architecture applicability of verified visual editing on InternVL3.5-8B.

A.3.2 ADDITIONAL DISCRIMINATIVE RESULTS ON AMBER

Table 6 summarizes reported AMBER-D results with LLaVA-1.5-7B and LLaVA-1.5-13B. For both LLaVA scales, VBEDIT uses the corresponding POPE-selected configuration without AMBER-specific parameter tuning. On LLaVA-1.5-7B, VBEDIT achieves 77.4% accuracy and 82.2% overall F1, the highest values listed in the table, exceeding the reported OPERA results by 1.4 and 0.8 points, respectively. On LLaVA-1.5-13B, the published VCD and DPO results lie in the 71–74%

Table 5: POPE results on InternVL3.5-8B. Matched Vanilla and VBEDIT adopt the same evaluation protocol, FINER-Tuning values are taken from its original paper. A dash denotes unreported F1.

Method	Random		Popular		Adversarial	
	Accuracy	F1	Accuracy	F1	Accuracy	F1
Matched Vanilla	92.58	92.37	89.13	88.92	87.53	87.49
FINER-Tuning (Xiao et al., 2026)	92.70	–	88.70	–	86.60	–
VBEDIT	**92.75**	**92.56**	**89.50**	**89.28**	**87.83**	**87.79**

range. Without AMBER-specific tuning, the POPE-selected configuration of VBEDIT attains 75.6% accuracy and 78.5% F1, exceeding the reported VCD results by 4.1 and 5.0 points, respectively. VBEDIT also exceeds the reported results of the training-based DPO baseline.

Table 6: AMBER-D comparison with published results on LLaVA-1.5-7B and 13B. † denotes a training-based method.

Method	Accuracy	Overall F1
LLaVA-1.5-7B		
VCD (Leng et al., 2024)	67.3	71.1
M3ID (Favero et al., 2024)	67.3	70.9
MoD (Chen et al., 2025)	72.4	75.3
OPERA (Huang et al., 2024)	76.0	81.4
VBEDIT	**77.4**	**82.2**
LLaVA-1.5-13B		
VCD (Leng et al., 2024)	71.5	73.5
DPO† (He et al., 2025)	71.7	73.7
VBEDIT	**75.6**	**78.5**

Table 7: Task-level AMBER-D results with LLaVA-1.5-7B and LLaVA-1.5-13B. Δ denotes the change from Matched Vanilla to VBEDIT.

Backbone	Method	Overall		Existence		Attribute		Relation	
		Acc.	F1	Acc.	F1	Acc.	F1	Acc.	F1
LLaVA-1.5-7B	Vanilla	76.5	80.5	77.9	87.5	77.1	73.9	**69.0**	**69.5**
	VBEDIT	**77.4**	**82.2**	**82.9**	**90.6**	**77.7**	**76.2**	59.4	65.7
	Δ	**+0.9**	**+1.7**	**+5.0**	**+3.1**	**+0.6**	**+2.3**	−9.6	−3.8
LLaVA-1.5-13B	Vanilla	70.9	72.6	67.0	80.2	73.1	65.5	**72.3**	53.1
	VBEDIT	**75.6**	**78.5**	**75.8**	**86.2**	**76.4**	**71.6**	71.9	**57.6**
	Δ	**+4.7**	**+5.9**	**+8.8**	**+6.0**	**+3.3**	**+6.1**	−0.4	**+4.5**

Table 7 provides task-level comparisons against the corresponding Matched Vanilla model. On LLaVA-1.5-7B, VBEDIT improves overall Accuracy and F1 by 0.9 and 1.7 points, respectively. The gains are concentrated on Existence and Attribute, whose F1 scores increase by 3.1 and 2.3 points, while Relation F1 decreases by 3.8 points. Relation questions often depend on joint evidence distributed across multiple objects, making some correct answers particularly sensitive to the counterfactual perturbation. Because R_{cf} measures decision sensitivity rather than correctness itself, the reliability gate can occasionally over-suppress such predictions. On LLaVA-1.5-13B, VBEDIT improves overall Accuracy and F1 by 4.7 and 5.9 points, with F1 gains on Existence, Attribute, and Relation. Relation Accuracy decreases only slightly by 0.4 points while its official F1 improves by 4.5 points, indicating that this behavior is substantially weaker and reflects a precision–recall trade-off rather than a broad failure on relation questions.

The lower absolute score of the 13B model is already present in its matched Vanilla baseline, which produces substantially more affirmative predictions than the 7B baseline. This non-monotonic scaling behavior is consistent with the original AMBER study, where LLaVA-1.5-13B achieves comparable Accuracy but lower F1 than its 7B counterpart on the discriminative task (Wang et al., 2023). We therefore interpret these results as within-model improvements rather than a controlled comparison of model scales.

A.3.3 CROSS-BENCHMARK EVALUATION ON HOPE

We further evaluate cross-benchmark transfer on the public description-based split of HOPE (Xie et al., 2025), which constructs challenging questions by pairing objects with image-inconsistent descriptions. The evaluation set contains 3,000 balanced Yes/No questions over 500 MS-COCO images. For each backbone, we use the same checkpoint, prompt, image processor, and greedy generated-text protocol as in the corresponding POPE experiment. For VBEDIT, the complete configuration—including the editing layer, attribution budget, counterfactual replacement strength, reliability threshold, and editing strength—is transferred from POPE without HOPE-specific parameter selection. VCD and HulluEdit likewise use their corresponding official POPE configurations. For VCD, we retain the official POPE hyperparameters, distorted-image construction, and contrastive scoring at both LLaVA scales, and apply deterministic argmax to match the common greedy protocol. HulluEdit is evaluated on LLaVA-1.5-7B using its released POPE configuration, and we do not report the 13B results since the official release does not specify a reproducible 13B POPE setting.

Table 8: Cross-benchmark transfer to the public HOPE description-based set with LLaVA-1.5-7B and 13B. All intervention methods use their POPE configurations without HOPE-specific parameter tuning. HulluEdit is reported only for LLaVA-1.5-7B since its released implementation specifies a 7B POPE configuration but no reproducible 13B counterpart. $\mathrm{F1_Y}$ and $\mathrm{F1_N}$ treat Yes and No as the positive class, respectively. Macro-F1 averages the two class-wise F1 scores.

Method	Accuracy	$\mathrm{F1_Y}$	$\mathrm{F1_N}$	Macro-F1
LLaVA-1.5-7B				
Vanilla	61.87	69.53	49.07	59.30
VCD (Leng et al., 2024)	60.80	69.20	46.10	57.65
HulluEdit (Lin et al., 2026)	62.20	69.82	49.42	59.62
VBEDIT	**67.80**	**71.30**	**63.33**	**67.31**
LLaVA-1.5-13B				
Vanilla	61.70	69.74	47.84	58.79
VCD (Leng et al., 2024)	63.43	70.74	51.27	61.00
VBEDIT	**71.13**	**72.94**	**69.07**	**71.00**

As shown in Table 8, on LLaVA-1.5-7B, HulluEdit provides a modest improvement over Vanilla, increasing Accuracy from 61.87% to 62.20% and Macro-F1 from 59.30% to 59.62%. VCD reduces Accuracy to 60.80% and Macro-F1 to 57.65%, indicating that its POPE configuration does not transfer favorably to the description-based questions under the matched greedy protocol. In comparison, VBEDIT reaches 67.80% Accuracy and 67.31% Macro-F1, exceeding HulluEdit by 5.60 and 7.69 points, respectively, and VCD by 7.00 and 9.66 points. These results indicate stronger cross-benchmark transfer of sample-specific verification under the matched 7B protocol.

On LLaVA-1.5-13B, POPE-configured VCD improves Accuracy from 61.70% to 63.43% and Macro-F1 from 58.79% to 61.00%. Under the same matched protocol, VBEDIT reaches 71.13% Accuracy and 71.00% Macro-F1, exceeding VCD by 7.70 and 10.00 points, respectively. It also improves both class-wise F1 scores over Vanilla: $\mathrm{F1_Y}$ increases from 69.74% to 72.94%, while $\mathrm{F1_N}$ increases from 47.84% to 69.07%. We omit HulluEdit from the 13B comparison since its released implementation does not specify an auditable 13B POPE configuration, and we avoid extrapolating the 7B settings without supporting evidence. Together, the two within-backbone comparisons show that the POPE-selected configuration transfers positively to HOPE at both LLaVA scales without HOPE-specific tuning; the different absolute gains should not be interpreted as a controlled model-scaling comparison.

A.4 Additional ablation results

This section supplements the main-paper ablations with mechanism diagnostics, replacement-target controls, reliability-gate randomization, global answer-bias control, and complete sensitivity analyses. Unless otherwise stated, all experiments use LLaVA-1.5-7B on the full POPE benchmark and retain the selected VBEdit configuration while changing only the factor under study.

A.4.1 Counterfactual replacement target

The complete method replaces attributed tokens with the within-image visual mean. Table 9 compares this choice with a zero vector and random visual tokens sampled from unselected locations in the same image. Random replacement results are reported over three predetermined seeds. The within-image mean yields the highest Accuracy, F1, and ΔCorrect. Although random replacement corrects slightly more false positives on average (270.7 versus 262), it also suppresses substantially more true positives (172.3 versus 153), reducing ΔCorrect from $+109$ to $+98.3$. Zero vector replacement shows a similar trade-off. These results support the within-image visual mean as a more effective image-conditioned counterfactual reference than arbitrary masking or replacement.

Table 9: Effect of the counterfactual replacement target on full POPE with LLaVA-1.5-7B. Random in-image replacement is averaged over three predetermined seeds.

Replacement	Accuracy	Precision	Recall	F1	FP→TN	TP→FN	ΔCorrect↑
Zero vector	87.51	88.96	85.93	87.42	265	168	+97
Random in-image	87.52	**89.06**	85.84	87.42	270.7	172.3	+98.3
Within-image visual mean	**87.64**	88.93	**86.27**	**87.58**	262	153	**+109**

A.4.2 Reliability-gate selectivity

The matched-random control in Table 3 isolates selection quality from the number of edits. We run five matched-random gates using exactly the same per-split intervention budget as VBEdit. Averaged over these five predetermined generated-text runs, matched-random gating obtains 86.67% Accuracy, 85.67% Precision, 88.39% Recall, and 87.01% F1. It produces 79.8 FP→TN transitions and 57.4 TP→FN transitions on average, corresponding to only 22.4 ΔCorrect, compared with $+109$ for VBEdit. These runs quantify performance under matched random selection.

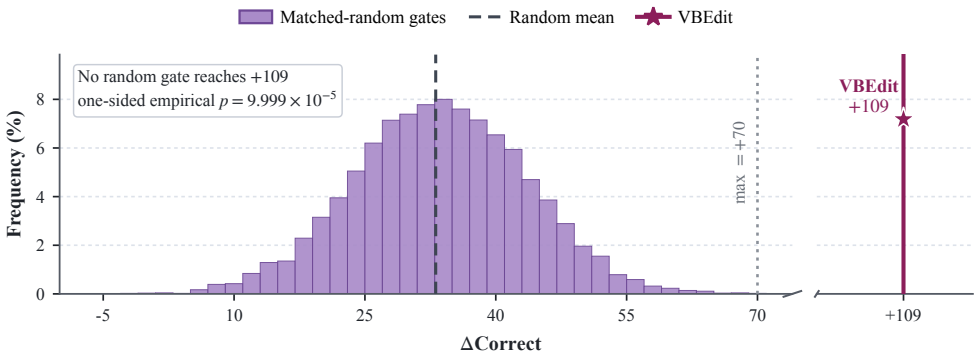

Figure 6: ΔCorrect distribution of 10,000 matched-random reliability gates on full POPE with LLaVA-1.5-7B. Each random gate uses the same per-split intervention budget as VBEdit. The dashed line marks the random-gate mean, and the star marks VBEdit. The broken horizontal axis omits an empty interval between the largest random outcome ($+70$) and VBEdit ($+109$). VBEdit exceeds every sampled random gate, giving a one-sided empirical p-value of 9.999×10^{-5}.

Figure 6 provides a larger matched-budget randomization analysis. Specifically, it shows that the 10,000 matched-budget Monte Carlo trials have mean 33.11, standard deviation 9.90, and maximum $+70$ in ΔCorrect. The $+109$ obtained by VBEdit exceeds every sampled random gate, yielding

a one-sided empirical p-value of 9.999×10^{-5}. Unlike the five generated-text runs, this larger experiment characterizes the variability of randomized selection over a fixed intervention budget. It therefore supports the selectivity of R_{cf}, but does not constitute an independent test-set generalization guarantee.

A.4.3 GLOBAL ANSWER-BIAS CONTROL

To examine whether VBEDIT's gains can be explained by uniformly suppressing affirmative answers, we introduce a global first-step answer-bias control. It adds the same scalar bias to all semantic Yes-token logits at the first generation step, without using visual attribution, counterfactual verification, sample-specific gating, or hidden-state editing. We consider two operating points: a deliberately strong in-benchmark bias selected to maximize POPE performance (**Accuracy-selected Yes bias**), and a label-free matched-flip bias selected only to approximate the number of semantic decisions changed by VBEDIT (**Flip-count-matched Yes bias**). These controls are intended to isolate the source of the improvement rather than to estimate held-out generalization.

All results are shown in Table 10. The accuracy-selected Yes bias improves matched Vanilla to 87.38% accuracy and 87.46% F1, but remains below VBEDIT at 87.64% and 87.58%, respectively. More importantly, under nearly identical intervention budgets (413 versus 415 semantic flips), the Flip-count-matched Yes bias yields only $+71$ net correct decisions, compared with $+109$ for VBEDIT. It corrects 242 false positives while suppressing 171 true positives, whereas VBEDIT corrects 262 and suppresses 153. Thus, uniform Yes suppression explains part of the aggregate improvement, but answer-conditioned verification identifies a more favorable set of interventions.

Table 10: Generated-text comparison with uniform first-step Yes-bias controls on full POPE. The accuracy-selected control maximizes overall POPE accuracy, whereas the flip-count-matched control is chosen without labels to approximate VBEDIT's number of semantic flips.

Method	Acc.	Prec.	F1	Flips	FP→TN	TP→FN	ΔCorrect
Matched Vanilla	86.42	84.41	86.96	0	0	0	0
Accuracy-selected Yes bias	87.38	87.85	87.46	320	203	117	+86
Flip-count-matched Yes bias	87.22	88.48	87.15	413	242	171	+71
VBEDIT	**87.64**	**88.93**	**87.58**	415	262	153	**+109**

Table 11: Sensitivity to the number of attributed visual tokens.

k	**Accuracy**	**Precision**	**Recall**	**F1**	**FP→TN**	**TP→FN**	**ΔCorrect**	$\mathbf{AUC_{FP>TP}}$
8	86.73	85.49	**88.80**	87.12	67	39	+28	75.41
16	87.32	87.53	87.33	87.43	185	105	+80	79.77
32	**87.64**	88.93	86.27	**87.58**	262	153	**+109**	83.16
64	87.52	89.20	85.67	87.40	278	180	+98	84.83
128	87.54	**89.24**	85.67	87.42	280	180	+100	85.74

Table 12: Sensitivity to counterfactual replacement strength α.

α	**Accuracy**	**F1**	**FP→TN**	**TP→FN**	**ΔCorrect**
0.000	86.42	86.96	0	0	0
0.125	86.53	87.00	31	21	+10
0.250	87.35	87.56	152	69	+83
0.375	87.26	87.28	216	141	+75
0.500	87.51	87.47	247	150	+97
0.625	87.58	87.54	250	147	+103
0.750	87.51	87.45	256	159	+97
0.875	87.59	87.52	260	156	+104
1.000	**87.64**	**87.58**	262	153	**+109**

Table 13: Sensitivity to the anchor layer.

Layer	Accuracy	Precision	Recall	F1	FP→TN	TP→FN	ΔCorrect
22	87.02	89.75	83.87	86.71	314	261	+53
24	86.49	89.77	82.67	86.07	321	315	+6
26	86.35	90.44	81.60	85.79	357	363	−6
28	86.63	**90.76**	81.87	86.09	370	351	+19
30	**87.64**	88.93	**86.27**	**87.58**	262	153	**+109**

A.4.4 Complete sensitivity results

Tables 11–13 provide the complete numerical results underlying Figure 3. Each setting is evaluated with complete inference while varying one structural hyperparameter and holding the remaining configuration fixed.

Table 11 shows that increasing k steadily improves the FP-over-TP AUC, but does not monotonically improve generated-text performance. Selecting $k = 32$ gives the best Accuracy, F1, and $\Delta\mathrm{Correct}$. Larger sets correct more false positives, but the additional exposure of true positives to editing increases harmful TP→FN transitions. Table 12 confirms exact recovery of Vanilla at $\alpha = 0$ and positive correction balances throughout the nonzero range. Moderate-to-strong replacement remains effective, with the selected $\alpha = 1$ attaining the highest Accuracy, F1, and $\Delta\mathrm{Correct} = +109$. The non-monotonic intermediate values indicate that replacement strength changes both gate activation and the realized answer-boundary crossings. Table 13 shows that the editing site is critical. Layers 22–28 correct many false positives but also incur 261–363 harmful TP→FN transitions. Layer 30 reduces this count to 153 and consequently achieves the best Accuracy, F1, and $\Delta\mathrm{Correct}$. Thus, strong token localization alone is insufficient unless the resulting direction is applied at an effective answer-level representation.

A.4.5 Threshold–strength landscape

The gate threshold and editing strength jointly control intervention coverage and the probability of crossing the answer boundary. We visualize their dense interaction using the cached affine candidate readout at the selected structural configuration. This analysis is explicitly reported as an offline candidate-probe landscape rather than generated-text performance; all main operating-point and one-dimensional sensitivity conclusions are confirmed with complete generated responses.

As illustrated in Figure 7, weak edits rarely cross the answer boundary, whereas overly broad or strong intervention increases the risk of suppressing correct positive predictions. The selected configuration lies in a high-performing region of the measured landscape rather than at an isolated numerical spike. Because the heatmap is constructed from the affine candidate readout, we use it only to characterize the local parameter landscape; the frozen operating point and the one-dimensional sensitivity trends are evaluated with full generated-text inference.

A.5 Computational overhead and efficiency

We measure generated-text inference on a fixed, stratified subset of 600 POPE examples (200 per split) with LLaVA-1.5-7B and batch size one. Vanilla and VBEdit use the same otherwise idle RTX 5090 and identical generation settings. Table 14 reports means across five runs after warm-up. Timing includes model inference through complete generation but excludes model loading, disk access, and input preprocessing.

On full POPE, VBEdit improves accuracy by 1.22 percentage points and F1 by 0.62 points, yielding 109 additional correct predictions relative to Vanilla (Table 3). On the efficiency subset, it adds 182.12 ms per sample ($1.8\times$ latency) and reduces throughput by 44.89%, while peak allocated memory increases by 36.7 MiB. The added computation is concentrated in evidence verification rather than the final edit. In separate stage profiling, signed attribution and its activation backward pass take 73.39 ms per sample, and counterfactual verification takes 70.27 ms; the gate decision and first-step hidden-state edit take only 1.39 and 1.70 ms, respectively. Since the gate depends on R_{cf}, gate-off examples still incur verification costs: the measured gate-on and gate-off latency ratios are $1.82\times$

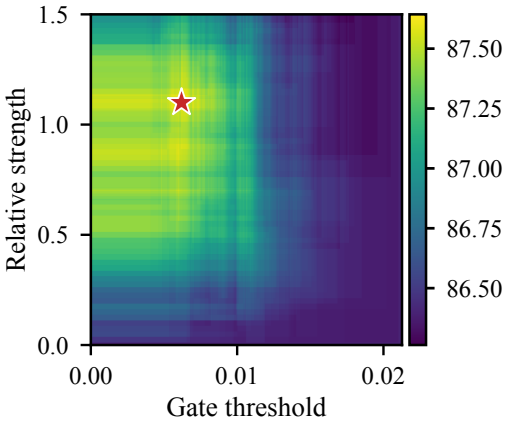

Figure 7: Offline candidate-probe Accuracy (%) over the counterfactual gate threshold and editing strength on full POPE. The star marks the frozen generated-text operating point.

Table 14: Performance and inference efficiency on POPE with LLaVA-1.5-7B. Accuracy and F1 use all 8,910 examples; efficiency uses a fixed 600-example subset averaged over five runs on one RTX 5090. The latency ratio is relative to Matched Vanilla; memory denotes peak allocated GPU memory.

Method	**Acc. (%)**	**F1 (%)**	**Latency (ms)**	**Ratio**	**Samples/s**	**Peak mem. (GiB)**
Matched Vanilla	86.42	86.96	223.66	1.0×	4.47	14.059
VBEDIT	87.64	87.58	405.78	1.8×	2.46	14.095

and 1.81×. The current implementation trades per-query latency for sample-specific verification and selective correction, with little additional peak memory.

A.6 QUALITATIVE ANALYSIS

This section provides representative sample-level analyses of VBEDIT's verification and editing behavior. We first compare different questions about the same image to illustrate that the counterfactual response is conditioned on the current answer rather than the image alone (Section A.6.1). We then present successful hallucination corrections and well-grounded predictions preserved by the reliability gate (Section A.6.2), followed by several failure cases (Section A.6.3). Each example reports the question, selected visual-token positions, counterfactual response, gate decision, and the corresponding Vanilla and VBEDIT outputs.

A.6.1 ANSWER-CONDITIONED VERIFICATION FOR DIFFERENT QUESTIONS ABOUT THE SAME IMAGE

Figures 8 and 9 compare different object questions about the same COCO image. Each page has one pair per POPE split: a high-response false positive is corrected, while a different grounded positive remains unchanged. These examples illustrate the answer-conditioned character of the verification signal. Translucent magenta cells project the selected top-32 visual-token positions onto the image grid. The actual counterfactual replacement and first-step edit act on hidden states.

A.6.2 CORRECTED HALLUCINATIONS AND PRESERVED CORRECT ANSWERS

Figure 10 shows nine visually reviewed FP → TN corrections, three from each split. Figure 11 shows six TP → TP answers preserved by the gate. Examples were selected from all 8,910 generated-text records and stratified by split and transition type.

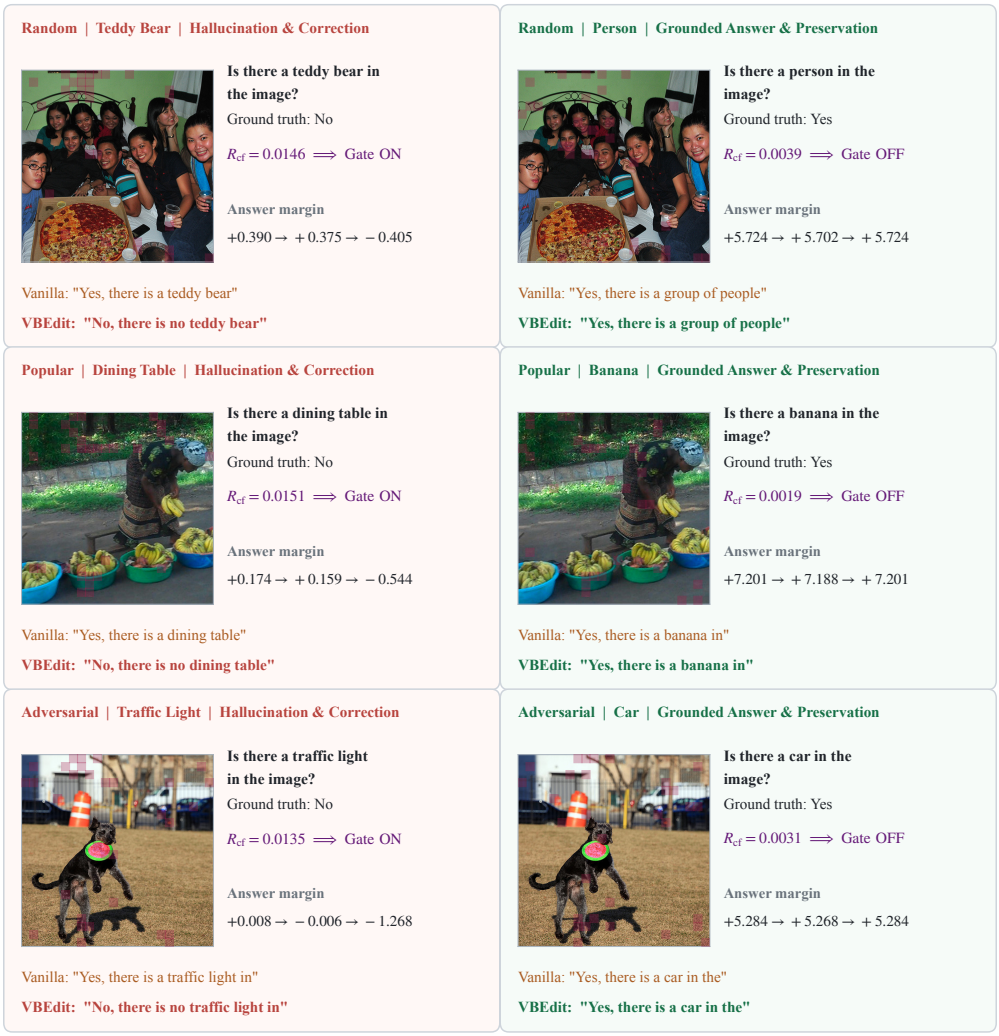

Figure 8: Answer-conditioned verification for different questions about the same image. Within each row, the left question receives a hallucinated *Yes* from Vanilla that VBEDIT changes to *No*; the right question concerns a different object in the same image and retains a correct *Yes*.

A.6.3 HARMFUL EDITS AND UNCORRECTED HALLUCINATIONS

Harmful edits. Figure 12 shows six correct *Yes* answers changed to incorrect *No* answers. For the Random surfboard example (top left), $R_{\text{cf}} = 0.0091$ activates the gate, and the answer margin changes from $+0.417$ to -0.101 after editing. The Adversarial cup example (bottom right) similarly changes from $+0.203$ to -1.036. Both examples illustrate a selection failure: the gate activates on correct answers, which editing then turns into errors. A counterfactual response can be large for a correct prediction since it measures the decision's sensitivity to the selected visual tokens, not the probability that the answer is hallucinated.

Uncorrected hallucinations. Figure 13 separates two ways a false positive can survive. In the Random teddy-bear example (top left), $R_{\text{cf}} = 0.0051$ leaves the gate off, and the incorrect *Yes* answer is unchanged. In the Random TV example (top right), the gate is on, but editing only reduces the positive margin from $+1.366$ to $+0.645$; the generated answer remains *Yes*. The Adversarial dining-table example (bottom right) exhibits the same pattern, ending at a positive margin of $+0.191$. These examples show why a false positive can remain uncorrected: either the gate stays off, or it turns on but editing still leaves the answer as *Yes*.

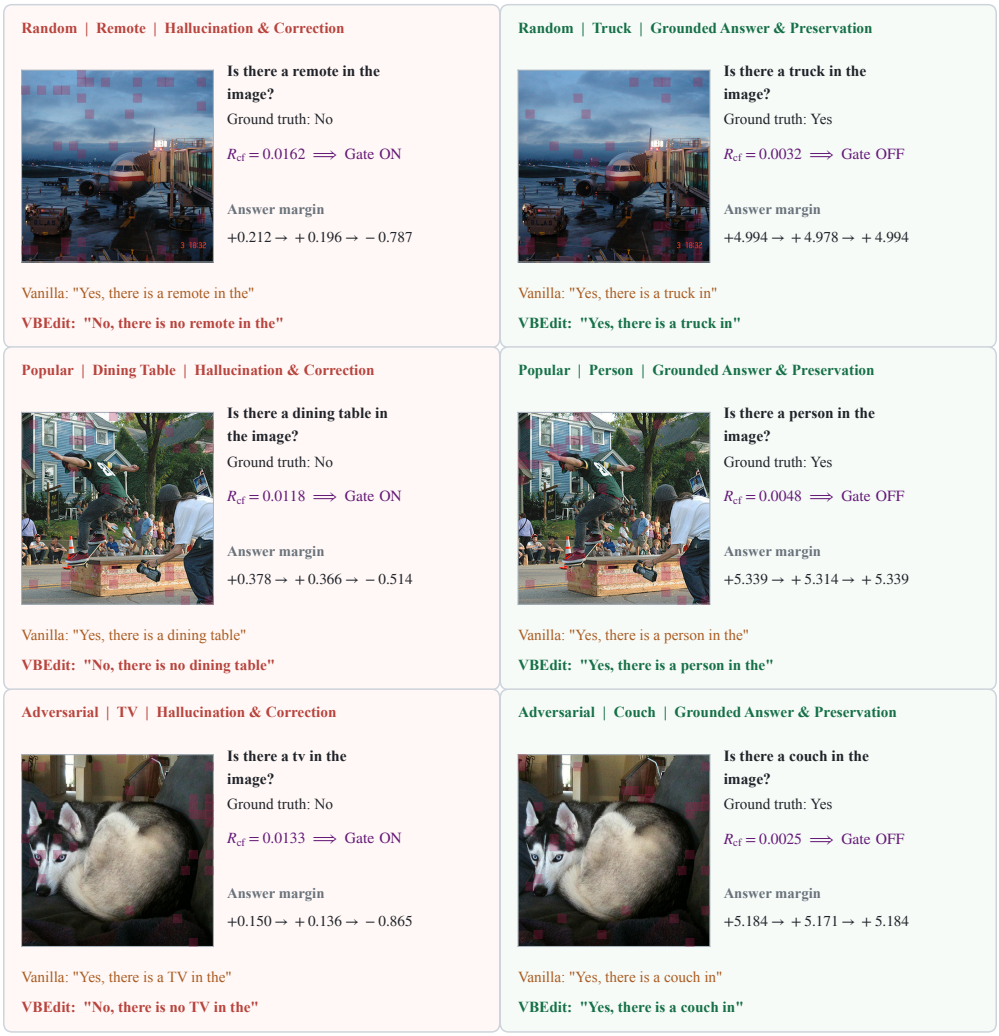

Figure 9: Additional examples of answer-conditioned verification for different questions about the same image. Within each row, the left question receives a hallucinated *Yes* from Vanilla that VBEDIT changes to *No*; the right question concerns a different object in the same image and retains a correct *Yes*.

A.7 LIMITATIONS AND FUTURE DIRECTIONS

Dependence on the base model's visual representation. As a training-free intervention on a frozen MLLM, VBEDIT can use only visual evidence that the model has encoded. When an object or attribute is not represented reliably, signed attribution and first-step editing cannot reconstruct that missing information. Improving the base model's visual representations, particularly for fine-grained details, is therefore complementary to verifying evidence before an edit.

Extension to broader answer spaces. VBEDIT seamlessly applies to binary visual queries by defining a semantic margin between competing answer candidates and verifying the relevant visual evidence before editing. Our current evaluation focuses on this setting, allowing the intervention behavior to be examined consistently across both answer classes. Extending the framework to multi-candidate tasks requires generalizing the semantic margin and counterfactual verification to multiple competing alternatives. Open-ended generation further requires constructing candidate sets dynamically and determining whether verification should extend beyond the first decoding step. We leave evaluation in these broader answer spaces for future work.

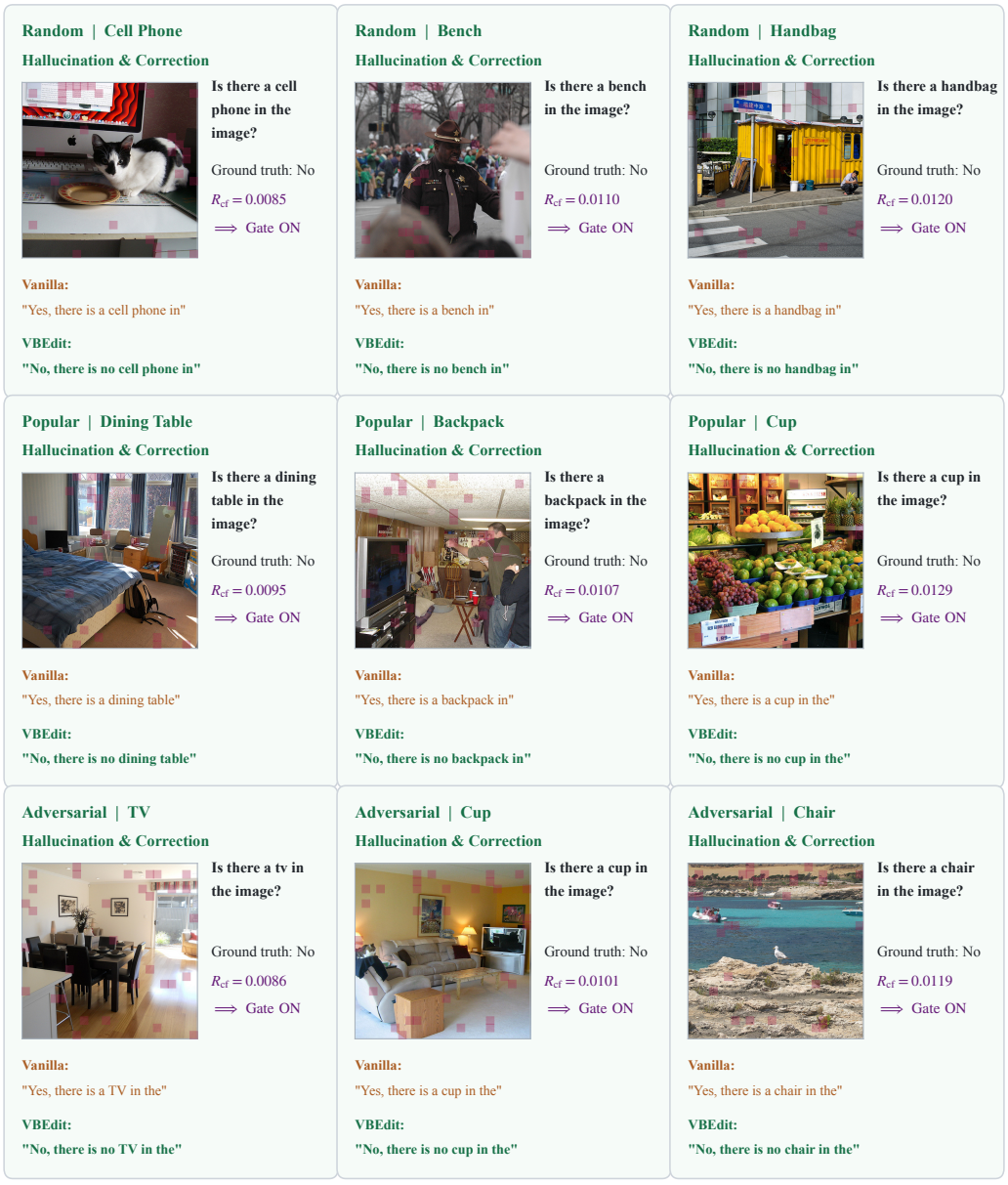

Figure 10: Nine generated-text false-positive corrections, arranged by POPE split.

Interpretation of representation-space counterfactuals. Our counterfactual probe perturbs contextualized visual-token states rather than image pixels or semantic objects. Accordingly, R_{cf} tests the current decision's sensitivity to an internal perturbation; it should not be interpreted as an image-level causal effect. Likewise, highlighted source-grid positions are neither object masks nor pixel-level causal localizations. Establishing a closer correspondence between these feature-level probes and controlled image-space interventions remains an open direction.

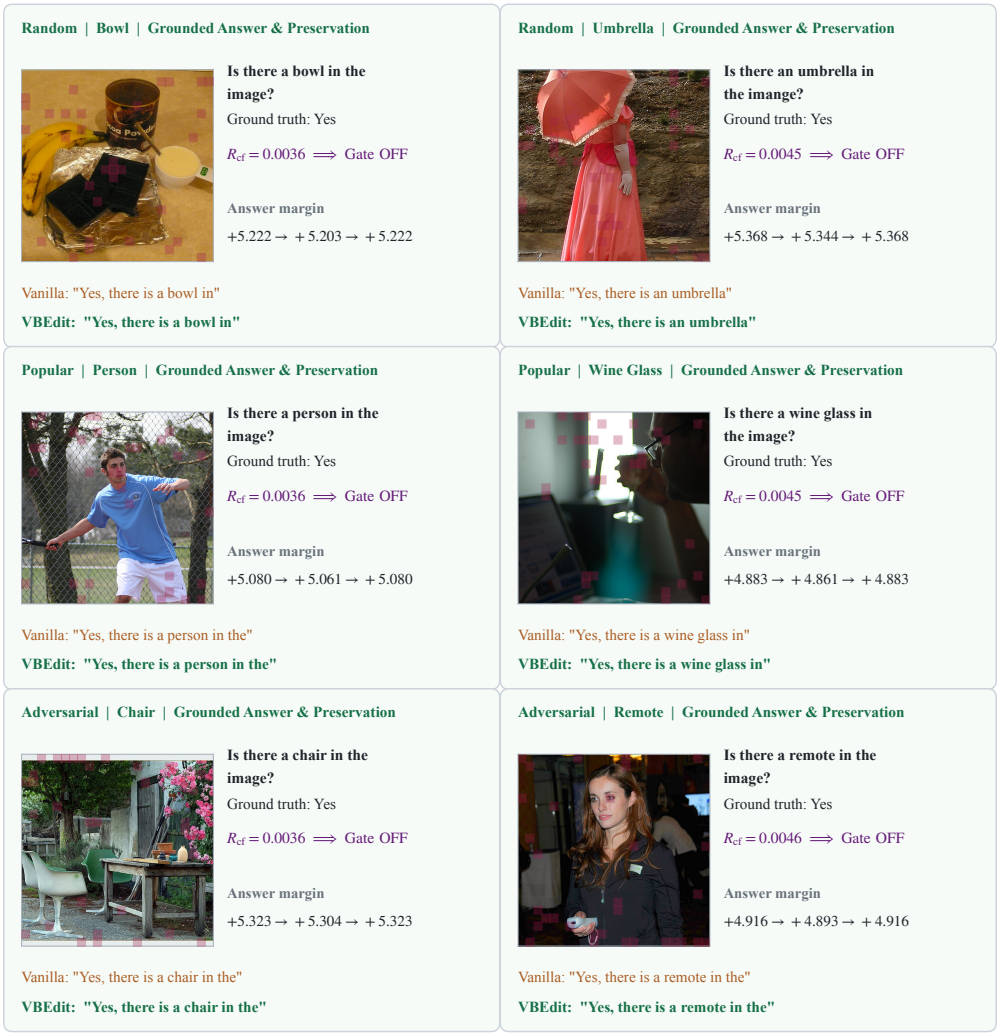

Figure 11: Six correct *Yes* predictions preserved by the gate, two per POPE split. Their counterfactual responses fall below the selected threshold, and the generated answers remain unchanged.

Random | Surfboard | Grounded Answer & Suppression



Is there a surfboard in the image?
Ground truth: Yes
$R_{cf} = 0.0091 \Longrightarrow$ Gate ON

Answer margin
$+0.417 \rightarrow +0.408 \rightarrow -0.101$

Vanilla: "Yes, there is a surfboard"
VBEdit: "No, there is no surfboard"

Random | Spoon | Grounded Answer & Suppression



Is there a spoon in the image?
Ground truth: Yes
$R_{cf} = 0.0119 \Longrightarrow$ Gate ON

Answer margin
$+0.137 \rightarrow +0.125 \rightarrow -0.677$

Vanilla: "Yes, there is a spoon in"
VBEdit: "No, there is no spoon in"

Popular | Potted Plant | Grounded Answer & Suppression



Is there a potted plant in the image?
Ground truth: Yes
$R_{cf} = 0.0097 \Longrightarrow$ Gate ON

Answer margin
$+0.270 \rightarrow +0.260 \rightarrow -0.258$

Vanilla: "Yes, there is a potted plant"
VBEdit: "No, there is no potted plant"

Popular | Parking Meter | Grounded Answer & Suppression



Is there a parking meter in the image?
Ground truth: Yes
$R_{cf} = 0.0122 \Longrightarrow$ Gate ON

Answer margin
$+0.643 \rightarrow +0.631 \rightarrow -0.729$

Vanilla: "Yes, there is a parking meter"
VBEdit: "No, there is no parking meter"

Adversarial | Baseball Glove | Grounded Answer & Suppression



Is there a baseball glove in the image?
Ground truth: Yes
$R_{cf} = 0.0096 \Longrightarrow$ Gate ON

Answer margin
$+0.991 \rightarrow +0.982 \rightarrow -0.111$

Vanilla: "Yes, there is a baseball glove"
VBEdit: "No, there is no baseball glove"

Adversarial | Cup | Grounded Answer & Suppression



Is there a cup in the image?
Ground truth: Yes
$R_{cf} = 0.0104 \Longrightarrow$ Gate ON

Answer margin
$+0.203 \rightarrow +0.193 \rightarrow -1.036$

Vanilla: "Yes, there is a cup in the"
VBEdit: "No, there is no cup in the"

Figure 12: Six harmful TP $\rightarrow$ FN edits, two per POPE split. The ground-truth label is *Yes*, but the selected first-step edit changes the parsed answer to *No*.

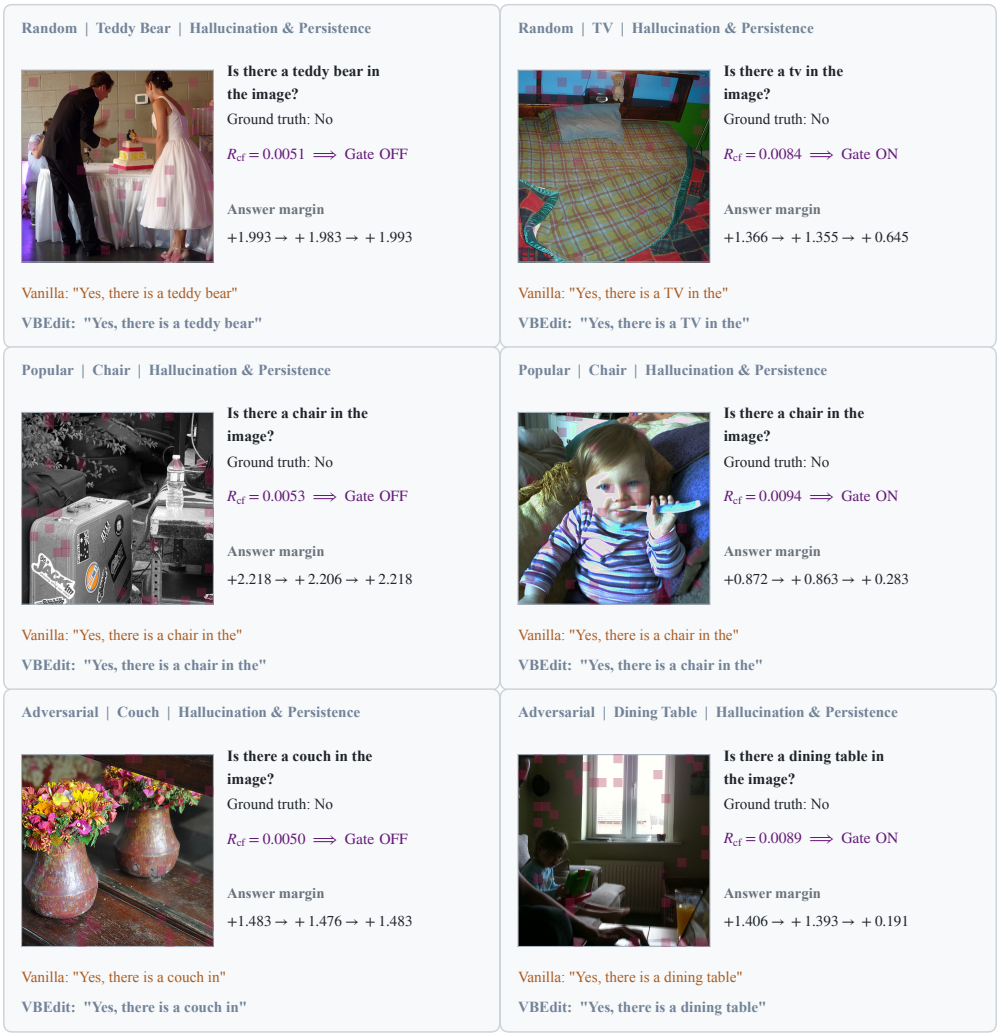

Figure 13: Six uncorrected false positives, two per POPE split. Within each split, one example is below the gate threshold; the other is above it but retains a parsed *Yes* answer. Neither gating nor editing guarantees correction.