Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Dual-Counterfactual Evidence Decomposition and Selective Editing for Hallucination Mitigation in Large Vision-Language Models

Zhiyong Jia
Ningxia University
Ziqiang Ma
maziqiang@nxu.edu.cn

Ningxia University
Jie Guo
Ningxia University

Research Article

Keywords:

Posted Date: November 2nd, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11118276/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Dual-Counterfactual Evidence Decomposition and Selective Editing for Hallucination Mitigation in Large Vision-Language Models

Zhiyong Jia[1], Ziqiang Ma[1*], Jie Guo[1]

[1] School of Information Engineering, Ningxia University, Yinchuan, 750021, China.

*Corresponding author(s). E-mail(s): maziqiang@nxu.edu.cn;
Contributing authors: 15969742062@163.com; 18795241649@163.com;

Abstract

Large vision-language models (LVLMs) often generate fluent but visually unsupported content. Existing training-free methods mainly recalibrate output distributions or edit representations without identifying whether the evidence behind the current claim is visual, linguistic, or shared. This limits intervention precision: unsupported language priors may survive, whereas grounded visual information may be weakened. We introduce a dual-counterfactual evidence decomposition and selective editing framework. Paired visual and linguistic counterfactuals act as sample-specific evidence probes, whose hidden-state responses are decomposed into shared, visual-exclusive, and linguistic-exclusive components. A claim-aware gate attenuates unsupported linguistic evidence only for visually verifiable claims, while an evidence-preserving update protects shared and visual-exclusive support. The method is training-free and requires neither parameter updates nor external verification models. Across three LVLM families, it achieves the highest Accuracy and F1 in all nine POPE backbone–split settings. On CHAIR, it obtains the lowest CHAIR_i on all three evaluated backbones, reducing it by 9.1%–14.7% relative to HulluEdit while maintaining comparable BLEU scores. It also achieves the highest aggregate MME-H on all three backbones, the best results on all four HallusionBench metrics, and a score of 29.3 on MM-Vet. At 23.59 tokens per second, the complete pipeline retains 75.5% of greedy-decoding throughput. These results establish evidence attribution before intervention as an effective and practical strategy for faithful LVLM generation.

Keywords: Large vision-language models · Hallucination mitigation · Counterfactual intervention · Multimodal grounding · Subspace decomposition · Representation editing

1 Introduction

Large vision-language models (LVLMs) have become a central interface for visual computing by combining pretrained visual encoders with the generation and reasoning capabilities of large language models [1–3]. They can describe complex scenes, answer visual questions, and perform multimodal reasoning through natural-language interaction. However, LVLMs still generate fluent content that is inconsistent with the input image, including nonexistent objects, incorrect attributes, unsupported quantities, and erroneous spatial relations [4, 5]. Because such hallucinations are often expressed confidently, their lack of visual support can be difficult to recognize from the response alone, undermining applications that require faithful visual scene understanding.

A major source of hallucination is the imbalance between visual evidence and linguistic priors during generation. When visual information is ambiguous or insufficiently preserved, statistical associations learned from large-scale text corpora can dominate next-token prediction. Training-based approaches address this problem through hallucination-aware instruction tuning, preference alignment, or correction data, but require additional supervision and parameter updates. Training-free approaches instead intervene during inference by contrasting distributions from original and perturbed inputs [6–8], recalibrating attention or visual signals [9–11], or suppressing hallucination-related representation directions [12–15]. Despite their progress, these methods mainly determine which prediction, signal, or direction should be adjusted. They do not explicitly determine whether the evidence for the current visually verifiable claim is shared across modalities, uniquely grounded in the image, or associated primarily with the linguistic context. Their interventions therefore rely largely on indirect estimates of hallucination risk rather than a sample-specific attribution of the underlying evidence.

This distinction matters because linguistic information is not inherently unreliable: it supports fluent expression, semantic composition, and descriptions of correctly perceived content. Conversely, activation induced by an image does not necessarily constitute sufficient evidence for a particular claim. A correct object description may be supported by both visual observations and compatible linguistic priors, whereas an unsupported object may arise predominantly from language-specific associations. Uniformly suppressing linguistic influence or amplifying visual activation cannot separate these cases and may disturb information that already has reliable visual support. Effective mitigation should therefore identify how the two modalities contribute to the current claim before deciding whether and what to edit.

We formulate this requirement as a selective evidence-attribution problem and introduce dual-counterfactual evidence probing. Under an aligned decoding context, a visual counterfactual weakens image evidence, whereas a linguistic counterfactual weakens textual evidence. The resulting hidden-state differences serve as sample-specific probes of the model's dependence on each modality. Rather than using the perturbed branches only to contrast output distributions, we decompose their paired responses into shared, visual-exclusive, and linguistic-exclusive evidence. This decomposition produces an evidence profile for the current sample and decoding step, distinguishing cross-modal support from modality-exclusive influence before intervention.

Based on this profile, our claim-aware selective editing mechanism intervenes only at decoding steps associated with visually verifiable content. When a claim is dominated by linguistic-exclusive evidence without sufficient visual support, the corresponding representation component is attenuated. An evidence-preserving update simultaneously protects shared and visual-exclusive components. The method therefore targets unsupported evidence without making the model uniformly conservative or suppressing reliable visual details. The complete procedure is training-free and requires neither parameter updates nor external verification models.

Extensive experiments across three LVLM families and six complementary benchmarks demonstrate that the proposed framework consistently reduces hallucinations while preserving fine-grained visual perception and general multimodal capability. Component ablations identify dual-counterfactual probing as the primary source of hallucination reduction, while representation-level measurements confirm that the update is concentrated on linguistic-exclusive evidence and largely preserves shared and visual-exclusive components. The complete pipeline also maintains practical end-to-end inference throughput. Together, these results support the effectiveness, selectivity, and practical applicability of the proposed framework.

The main contributions are summarized as follows:

1. We formulate training-free hallucination mitigation as a sample-specific evidence-attribution problem and introduce paired visual and linguistic counterfactual probes that reveal how each modality supports the current claim.
2. We develop a shared–exclusive evidence decomposition and claim-aware selective-editing strategy that distinguishes aligned cross-modal support from modality-exclusive influence, attenuating unsupported linguistic evidence while preserving shared and visual-exclusive information.
3. We demonstrate consistent hallucination reduction across multiple LVLM families and evaluation regimes, together with preserved visual capability, mechanism-aligned representation changes, and practical end-to-end throughput.

2 Related Work

Hallucination mitigation in large vision-language models (LVLMs) follows two broad routes. Training-based methods improve visual grounding through additional supervision and parameter updates, whereas training-free methods modify the generation process of a frozen model. Because our framework belongs to the latter setting, we review the two inference-time directions most relevant to its design: decoding-based calibration, which changes how candidate tokens are scored, and representation-level intervention, which changes the internal evidence from which those scores are produced.

2.1 Decoding-Based Hallucination Mitigation

Decoding-based methods improve visual faithfulness by modifying token scores or generation policies without updating model parameters. Visual Contrastive Decoding (VCD) contrasts predictions from the original image with those from a globally

distorted image and discounts tokens that remain likely after visual evidence is degraded [6]. OPERA penalizes excessive reliance on a small set of summary tokens and reallocates beam-search attention through retrospection [9], whereas HALC combines local focal grounding with global decoding to correct object hallucinations using fine-grained visual context [7]. These methods establish output calibration as an effective training-free route, but the intervention is applied after multimodal evidence has already been compressed into token probabilities or a decoding policy.

Later work makes calibration increasingly dependent on the current input and decoding step. Octopus selects among multiple contrastive strategies according to the hallucination characteristics of the current input and generation state [8]. ALD^2 denoises visual tokens at a dynamically selected shallow layer and combines the resulting distribution with the final prediction [16]. Cross-Modal Attention Calibration (CMAC) selectively masks cross-modal value vectors and calibrates visual-token positions to reduce positional bias and spurious inter-modal correlations [10]. These developments make calibration more adaptive with respect to when and where it is applied. However, the resulting correction remains driven primarily by changes in output scores, attention patterns, or visually perturbed signals, rather than by an explicit attribution of the evidence supporting the current claim.

Counterfactual decoding is especially relevant to our formulation. VCD constructs a global visual counterfactual through image distortion, while CounterfactualLVLM masks influential objects identified by an auxiliary visual question answering model and contrasts the resulting distribution with the original prediction [17]. In both cases, the counterfactual branch serves mainly as a negative output reference. Our framework instead pairs visual and linguistic counterfactuals as internal evidence probes under the same decoding context. Their hidden-state responses are decomposed into shared, visual-exclusive, and linguistic-exclusive components before any edit is applied. Output contrast can identify predictions that change under perturbation, but it does not by itself determine whether the affected information reflects reliable cross-modal support or unsupported modality-exclusive influence. Evidence attribution before intervention therefore enables the claim-aware and evidence-preserving edit developed in this work.

2.2 Representation-Level Hallucination Mitigation

Representation-level methods intervene before internal states are projected onto output logits, enabling hallucination-related information to be modified within the model's activation space. Nullu derives a low-dimensional HalluSpace from representation differences between truthful and hallucinated samples and orthogonalizes model weights against the identified subspace [12]. The Hidden Life of Tokens traces token representations across transformer layers, revealing that visually grounded information may emerge at intermediate layers but weaken before the final prediction [13]. Its VISTA framework therefore reinforces informative early-layer signals during decoding. Together, these works show that intermediate representations expose structured signals that can be suppressed, recovered, or redirected without retraining the LVLM.

Subsequent work makes internal intervention more selective and sample-dependent. DiVE identifies visually informative layers and separates aggregated visual evidence from the mixed representation, then suppresses that component to construct a

language-prior-dominated reference distribution for contrastive decoding [14]. HulluEdit decomposes hidden states into orthogonal visual-evidence, conflicting-prior, and residual components and selectively suppresses the conflicting-prior subspace while protecting the estimated visual component [15]. This progression moves representation editing beyond global suppression by recognizing that language-related activation can carry valid multimodal information that should not be removed indiscriminately.

The key unresolved issue is not whether internal representations can be edited, but how the evidence behind the current claim should be attributed before editing. Existing methods derive directions from truthful–hallucinated contrasts, trace layer-wise visual information, or estimate visual and prior-related components from a particular internal signal. These approaches identify useful intervention targets, yet they do not explicitly compare matched visual and linguistic interventions under the same decoding context. Modality association alone therefore does not reveal whether a response is jointly supported by both modalities or is exclusive to one of them. This distinction matters because a visually correct claim may agree with linguistic knowledge, whereas linguistic-exclusive evidence becomes risky only when it lacks corresponding support from the image.

Our framework directly addresses this attribution gap with paired modality-specific counterfactual probes. Under an aligned decoding prefix, the visual and linguistic branches selectively weaken their respective modalities, and the resulting hidden-state responses are used to construct shared, visual-exclusive, and linguistic-exclusive evidence for the current generation step. The framework therefore separates evidence diagnosis from intervention: counterfactual decomposition determines which evidence is shared or exclusive, whereas the claim-aware gate determines whether an edit should be applied. This formulation advances representation editing from direct direction suppression to sample-specific evidence attribution followed by selective, evidence-preserving intervention.

3 Method

3.1 Problem Formulation and Framework Overview

Given an input image I and a textual instruction X, an LVLM autoregressively generates a response $Y = \{y_1, \ldots, y_T\}$. At decoding step t, standard LVLM decoding maps the final-layer hidden representation to a next-token distribution:

$$p(y_t \mid I, X, Y_{<t}) = \mathrm{Softmax}\left(W_o h_t^L\right), \tag{1}$$

where $h_t^L \in \mathbb{R}^d$ denotes the hidden representation at the final transformer layer, W_o is the output projection matrix, and $Y_{<t}$ contains the previously generated tokens. Because h_t^L is jointly conditioned on the image and linguistic context, it entangles the contributions of both modalities. The resulting distribution therefore does not reveal whether the current claim is supported by the image, the linguistic context, or both. This ambiguity is consequential: editing a language-sensitive direction without first attributing its evidence can remove reliably grounded information together with hallucination-related activation.

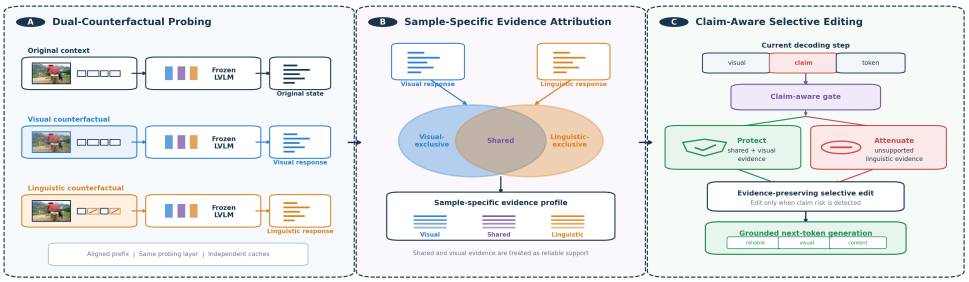


Fig. 1 Overview of the proposed dual-counterfactual evidence decomposition and selective editing framework

We address this ambiguity by formulating hallucination mitigation as evidence attribution before selective intervention. For the claim considered at decoding step t, we characterize its evidence profile as

$$\mathcal{E} = \{\mathcal{E}_s, \mathcal{E}_v, \mathcal{E}_l\}, \tag{2}$$

where $\mathcal{E}_s$, $\mathcal{E}_v$, and $\mathcal{E}_l$ denote shared, visual-exclusive, and linguistic-exclusive evidence, respectively. Crucially, linguistic-exclusive evidence is not itself a hallucination; it becomes potentially unreliable only when it dominates a visually verifiable claim without corresponding support from the image. The objective is therefore not to suppress linguistic influence globally, but to selectively attenuate unsupported linguistic-exclusive evidence while retaining shared and visual-exclusive components.

As illustrated in Fig. 1, the framework follows three stages. First, paired visual and linguistic counterfactuals probe how the current hidden representation depends on each modality. Second, their responses are decomposed into shared, visual-exclusive, and linguistic-exclusive evidence, producing a sample-specific evidence profile. Third, the claim-aware gate activates editing only for visually verifiable content. Once activated, selective editing attenuates unsupported linguistic-exclusive evidence, while the evidence-preserving update protects shared and visual-exclusive components. This ordering separates what evidence supports the current claim from when an intervention is required. The edited representation is then used for next-token prediction. The entire framework operates at inference time and requires neither additional LVLM training nor parameter updates.

3.2 Dual-Counterfactual Evidence Probing

The evidence profile in Eq. (2) cannot be recovered directly from a fused hidden state. We therefore estimate modality dependence through paired counterfactual interventions performed under an otherwise identical decoding context. The visual branch weakens image evidence, whereas the linguistic branch weakens textual evidence; their representation changes reveal how strongly the current state depends on each modality. Unlike counterfactual decoding, which turns perturbed inputs into negative output

distributions, these branches function only as internal evidence probes. Their outputs are not decoded as competing responses.

At decoding step t, let $Z_t = (X, Y_{<t})$ denote the linguistic context and $E_t = \mathrm{Emb}(Z_t)$ its input embeddings. The unmodified hidden state at the selected evidence-probing layer ℓ is

$$h_t^{o,\ell} = F^{\ell}(I, E_t) \in \mathbb{R}^d, \tag{3}$$

where $F^{\ell}(\cdot)$ denotes the LVLM computation up to layer ℓ. This original state provides a common reference for both counterfactual probes.

To measure dependence on visual evidence, we apply diffusion-style Gaussian perturbation to the normalized image tensor:

$$\widetilde{I} = \sqrt{\alpha_v} I + \sqrt{1-\alpha_v}\epsilon_v, \qquad h_t^{v,\ell} = F^{\ell}(\widetilde{I}, E_t), \quad \epsilon_v \sim \mathcal{N}(0, \mathbf{I}), \tag{4}$$

where α_v controls the amount of visual information retained in the counterfactual input. The perturbed image is sampled once and reused throughout generation, while the linguistic context remains identical to that of the original branch. Reusing the same counterfactual image prevents the probe itself from changing across autoregressive steps, so that each response is measured against a consistent weakened visual input.

To construct the complementary linguistic probe without changing token alignment, Gaussian perturbation is applied only to non-special text-token embeddings:

$$\widetilde{E}_t = (1 - M_t) \odot E_t + M_t \odot \left(\sqrt{\alpha_l} E_t + \sqrt{1-\alpha_l}\epsilon_l\right), \qquad h_t^{l,\ell} = F^{\ell}(I, \widetilde{E}_t), \tag{5}$$

where $\epsilon_l \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$, and M_t is a binary mask that equals one for ordinary linguistic tokens and zero for protected special tokens. The perturbation therefore changes ordinary text-token embeddings while preserving the special tokens, token sequence, sequence length, positional indices, and attention mask of the original branch. Because this branch is used only as an internal evidence probe, the perturbed embeddings need not correspond to discrete vocabulary items or form a fluent counterfactual sentence. Its purpose is not to generate a plausible alternative response, but to measure how the internal representation changes when linguistic evidence is weakened under an aligned context.

Gaussian perturbation provides a continuous intervention strength, whereas token masking and embedding dropout rely on discrete replacement or complete erasure. A coefficient close to one makes a counterfactual nearly identical to the original input and produces a weak response, whereas an excessively small coefficient can destroy the semantic structure required for a meaningful comparison. We use $\alpha_v = 0.8$ and $\alpha_l = 0.9$. The larger linguistic coefficient retains more of the original embedding information and therefore preserves the aligned linguistic structure more conservatively. Both coefficients remain fixed across samples, backbones, and evaluation datasets.

The modality-specific response at the current decoding position is defined as the normalized signed difference between the original and counterfactual hidden states:

$$r_t^{m,\ell} = \frac{h_t^{o,\ell} - h_t^{m,\ell}}{\left\| h_t^{o,\ell} - h_t^{m,\ell} \right\|_2 + \varepsilon}, \qquad m \in \{v, l\}, \qquad \mathcal{R}_t^{\ell} = \{r_t^{v,\ell}, r_t^{l,\ell}\}, \tag{6}$$

where ε ensures numerical stability. The signed difference encodes the representation direction induced by weakening the corresponding modality. L_2 normalization makes the visual and linguistic responses comparable by direction and prevents their raw magnitudes from dominating the subsequent positive-alignment decomposition. Responses are extracted from the current decoding position, which directly supplies the representation used for next-token prediction.

The original, visual-counterfactual, and linguistic-counterfactual branches use the same discrete generation prefix while maintaining independent key–value caches. Their autoregressive histories therefore remain token-wise aligned, allowing the observed differences to be attributed to the designated modality interventions rather than to mismatched generation histories. Crucially, $\mathcal{R}_t^{\ell}$ is not used as a direct logit correction. It is a diagnostic response pair passed to the shared–exclusive decomposition in the following section. This change in role—from a negative output reference to an internal evidence probe—enables the framework to identify shared and modality-exclusive evidence before applying any edit.

3.3 Shared–Exclusive Evidence Decomposition

The counterfactual responses $r_t^{v,\ell}$ and $r_t^{l,\ell}$ measure how the current representation changes when visual and linguistic information is weakened, respectively. They cannot, however, be interpreted directly as independent modality components: content supported jointly by both modalities may produce aligned changes in both branches. We therefore decompose the directional agreement between the two responses before interpreting their remaining components. For notational simplicity, we omit the probing-layer superscript ℓ in the remainder of this section and write the responses as r_t^v and r_t^l.

We define their directional agreement and the resulting shared evidence vector as

$$\begin{aligned} c_t &= (r_t^v)^\top r_t^l, \qquad \gamma_t = [c_t]_+ = \max(c_t, 0), \\ u_t^s &= \begin{cases} \dfrac{r_t^v + r_t^l}{\left\| r_t^v + r_t^l \right\|_2 + \varepsilon}, & c_t > 0 \text{ and } \left\| r_t^v + r_t^l \right\|_2 > \varepsilon, \\ \mathbf{0}, & \text{otherwise}, \end{cases} \\ e_t^s &= \frac{\gamma_t}{2} \left(\left| (u_t^s)^\top r_t^v \right| + \left| (u_t^s)^\top r_t^l \right| \right) u_t^s. \end{aligned} \tag{7}$$

Here, c_t measures the directional agreement between the normalized visual and linguistic responses, and γ_t retains only positive agreement. When the two responses are positively aligned and their sum is numerically non-degenerate, u_t^s gives their normalized consensus direction. The shared vector e_t^s scales this direction according to its support from both responses. This construction assigns shared evidence only to mutually aligned response directions. If the responses are conflicting, orthogonal, or degenerate, we set $u_t^s = e_t^s = \mathbf{0}$. The boundary condition follows directly from the positive-alignment rule because $\gamma_t = 0$ whenever $c_t \leq 0$.

After isolating the shared component, we obtain the visual-exclusive and linguistic-exclusive components as the corresponding residuals:

$$e_t^v = r_t^v - e_t^s, \qquad e_t^l = r_t^l - e_t^s. \tag{8}$$

The term "exclusive" has an operational meaning: it denotes the portion of a counterfactual response that is not explained by positive agreement between the two probes. It does not imply complete statistical or causal independence between modalities, nor does it mean that every exclusive component is erroneous.

The instantaneous components describe the evidence response at the current decoding step. To construct more stable sample-specific evidence subspaces, we accumulate the decomposed responses produced during the current generation and retain their leading directions. For each evidence type $k \in \{s, v, l\}$, we define

$$\begin{aligned} B_t^k &= [e_1^k, \ldots, e_t^k] = U_t^k \Sigma_t^k (V_t^k)^\top, \\ \widehat{r}_t^k &= \min\left\{r, \operatorname{rank}(B_t^k)\right\}, \\ P_t^k &= U_{t,\widehat{r}_t^k}^k \left(U_{t,\widehat{r}_t^k}^k\right)^\top, \qquad k \in \{s, v, l\}. \end{aligned} \tag{9}$$

Here, B_t^k collects the evidence responses observed for the current sample, r is the maximum retained rank, and $\widehat{r}_t^k$ is the effective rank supported by the available non-degenerate singular directions. The columns of $U_{t,\widehat{r}_t^k}^k$ are the corresponding leading left singular vectors, and P_t^k projects a representation onto the resulting evidence subspace.

During early decoding, the number of available response directions can be smaller than r. In particular, when $t < r$, the rank of B_t^k cannot exceed the number of accumulated responses. The effective-rank definition therefore retains only $\min\{r, \operatorname{rank}(B_t^k)\}$ valid directions instead of introducing undefined or zero-padded basis vectors. If B_t^k contains no valid singular direction, the corresponding projector is defined as $P_t^k = \mathbf{0}$. This rule allows each evidence basis to grow progressively as new responses become available. Because the subspaces are reconstructed from the current generation, they adapt to the image, instruction, and generated context rather than serving as globally fixed editing directions.

Finally, we summarize the instantaneous evidence composition at decoding step t as

$$\eta_t^k = \|e_t^k\|_2, \qquad \pi_t^k = \frac{\eta_t^k}{\eta_t^s + \eta_t^v + \eta_t^l + \varepsilon}, \qquad \Pi_t = (\pi_t^s, \pi_t^v, \pi_t^l). \tag{10}$$

Here, η_t^k measures the magnitude of evidence component k, whereas π_t^k expresses its relative contribution to the current response. A large linguistic-exclusive ratio π_t^l indicates that the linguistic response contains a substantial component that is not matched by positive visual–linguistic agreement. This ratio is a diagnostic signal rather than a hallucination decision: linguistic information may remain necessary for fluent and semantically coherent generation. The claim-aware mechanism introduced next therefore combines Π_t with the semantic status of the current prediction before determining whether selective attenuation should be activated.

3.4 Claim-Aware Selective Editing

A large linguistic-exclusive response is informative but not sufficient to identify a hallucination. Linguistic information also supports syntax, fluency, and semantic composition, and many generated tokens do not express content that can be verified directly from the image. The editing decision must therefore answer two questions jointly: whether the next-token prediction expresses a visual claim and whether that claim relies strongly on linguistic-exclusive evidence.

To identify visually verifiable predictions without introducing an auxiliary semantic model, we construct a benchmark-independent lexicon $\mathcal{L}_{\text{claim}}$. It covers five semantic categories: object nouns, visually observable attributes, quantities and numerals, spatial-relation terms, and observable action verbs. The lexicon is constructed without using image annotations, question labels, or answer labels from POPE, CHAIR, MME, AMBER, HallusionBench, or MM-Vet. For each LVLM backbone, its entries are mapped through the corresponding tokenizer to obtain a tokenizer-compatible set V_{claim}. The hard trigger retains only concepts represented by a single ordinary token and excludes special tokens and incomplete subword fragments. The semantic lexicon is therefore shared across backbones, while the associated token IDs remain tokenizer-specific. Once constructed, V_{claim} is fixed across images, prompts, decoding steps, and evaluation datasets.

Given the next-token distribution of the original branch, we measure the probability mass assigned to visually verifiable concepts as

$$p_t^o = \text{Softmax}(W_o h_t^{o,L}), \qquad q_t = \sum_{w \in V_{\text{claim}}} p_t^o(w). \tag{11}$$

The score q_t estimates the likelihood that the current decoding step will produce a token-level visual claim. It is not a sentence-level grounding score. Restricting the sum to single-token concepts makes the trigger deterministic and computationally lightweight, with no additional classifier or image annotation. Multi-token expressions are not included in the hard vocabulary sum because the probability of an isolated subword at one decoding step does not reliably determine the visual meaning of the complete expression.

We then combine the visual-claim likelihood q_t with the linguistic-exclusive evidence ratio π_t^l from Section 3.3. The gate and adaptive editing strength are defined as

$$g_t = \mathbb{I}(q_t \geq \tau), \qquad \rho_t = g_t q_t \pi_t^l, \qquad \lambda_t = \lambda_{\max} \rho_t, \tag{12}$$

where τ is the claim threshold and $\lambda_{\max}$ is the maximum editing strength. If $q_t < \tau$, then $g_t = 0$, so no representation update is applied. Once the claim gate is activated, the strength increases only when the prediction has both a high visual-claim likelihood and a large linguistic-exclusive evidence ratio. The first factor determines when visual verification is relevant, while the second determines how strongly the prediction depends on unmatched linguistic evidence. Their product therefore prevents either semantic claim likelihood or linguistic dependence alone from triggering a strong intervention.

Once the claim gate is activated, the linguistic-exclusive projector P_t^l, estimated at the evidence-probing layer ℓ, identifies the evidence direction to be attenuated. We apply this sample-specific projector to the final-layer representation of the original branch and form a preliminary edited state:

$$z_t^l = P_t^l h_t^{o,L}, \qquad \widehat{h}_t^L = h_t^{o,L} - \lambda_t z_t^l. \tag{13}$$

Equation (13) separates the layer used for evidence diagnosis from the layer used for representation intervention. The projector P_t^l is constructed from the counterfactual responses at layer ℓ, whereas the candidate edit is applied to the final-layer state $h_t^{o,L}$. The editing strength λ_t remains conditioned on the current visual-claim likelihood and evidence composition. When $g_t = 0$, we have $\lambda_t = 0$, and the original representation remains unchanged. When the claim gate is activated, the update becomes stronger as the predicted visual claim relies more heavily on linguistic-exclusive evidence.

The state $\widehat{h}_t^L$ remains preliminary because the sample-specific evidence subspaces need not be mutually orthogonal. A direction categorized as linguistic-exclusive may overlap with shared or visual-exclusive evidence that remains useful for visually grounded generation. Section 3.5 removes this protected overlap from the candidate edit before constructing the final representation update.

3.5 Evidence-Preserving Representation Update

The preliminary update in Eq. (13) targets linguistic-exclusive evidence, while reliable shared and visual-exclusive directions should remain protected. Because the sample-specific evidence subspaces are estimated from interacting multimodal representations, they may partially overlap. Directly subtracting z_t^l could therefore alter visually grounded information carried by the protected evidence. We address this issue by constructing a protected subspace from the shared and visual-exclusive bases and filtering the candidate edit against this subspace.

Using the low-rank bases defined in Section 3.3, we construct an orthonormal basis for the protected evidence and its corresponding projector:

$$U_t^p = \mathrm{Orth}\left(\left[U_{t,\widehat{r}_t^s}^s, U_{t,\widehat{r}_t^v}^v\right]\right), \qquad P_t^p = U_t^p (U_t^p)^\top. \tag{14}$$

Here, $\mathrm{Orth}(\cdot)$ removes linearly dependent directions and returns an orthonormal basis spanning the union of the shared and visual-exclusive evidence subspaces. Together, these subspaces represent the two sources of evidence that the update is designed to preserve: positive cross-modal agreement and visual-exclusive support. We then remove the protected component from the candidate linguistic edit:

$$d_t^l = (I_d - P_t^p) z_t^l, \qquad \kappa_t = \frac{\|d_t^l\|_2^2}{\|z_t^l\|_2^2 + \varepsilon}. \tag{15}$$

The residual d_t^l is the portion of the candidate edit that lies outside the protected evidence subspace. The ratio $\kappa_t \in [0, 1]$ measures how much of the candidate editing energy remains after evidence protection. A value close to one indicates that the linguistic-exclusive direction is largely separable from the protected evidence, whereas a small value indicates substantial overlap. This ratio allows the final update

to respond to the actual separability of the evidence rather than enforcing the same amount of subtraction in every case.

The preservation coefficient $\beta \in [0, 1]$ determines how strongly the protected-evidence filtering influences the final edit. We interpolate between the original linguistic-exclusive component z_t^l and its protection-filtered residual d_t^l. The final edit vector, updated hidden state, and next-token distribution are defined as

$$\begin{aligned} \bar{z}_t^l &= (1-\beta) z_t^l + \beta \kappa_t d_t^l, \\ h_t^{*,L} &= h_t^{o,L} - \lambda_t \bar{z}_t^l, \\ p_t^* &= \mathrm{Softmax}\left(W_o h_t^{*,L}\right). \end{aligned} \tag{16}$$

When $\beta = 0$, Eq. (16) reduces to the direct claim-aware update introduced in Section 3.4. As β increases, the edit shifts toward the protection-filtered residual. At $\beta = 1$, only the separable component d_t^l contributes, and its magnitude is further modulated by κ_t. Consequently, strong overlap with shared or visual-exclusive evidence produces a more conservative update, whereas a well-separated linguistic-exclusive direction can be attenuated more fully.

Together, Eqs. (12)–(16) implement three levels of selectivity. The claim gate g_t determines when editing is activated; the linguistic-exclusive and protected projectors determine which evidence directions are modified or retained; and λ_t, κ_t, and β determine how strongly the final-layer representation is updated. The resulting state $h_t^{*,L}$ is mapped directly to the output logits. The next token is selected from p_t^* using the original decoding strategy and appended to the prefix shared by all three branches.

3.6 Overall Inference Procedure and Complexity

The proposed method augments standard autoregressive decoding with sample-specific evidence probing and selective final-layer editing, without retraining the LVLM. The visual counterfactual is constructed once for each image and reused throughout generation. At decoding step t, the original, visual-counterfactual, and linguistic-counterfactual branches consume the same discrete prefix. The original branch passes through all L transformer layers to produce the final-layer representation $h_t^{o,L}$ and the preliminary next-token distribution required by the claim gate. The two counterfactual branches terminate at the evidence-probing layer ℓ, where their responses are used to construct the shared and modality-exclusive evidence subspaces.

The resulting evidence projectors diagnose the contributions of the visual input and linguistic context at layer ℓ. The claim-aware gate and evidence-preserving mechanism then use these projectors to edit the final-layer representation $h_t^{o,L}$. The edited state $h_t^{*,L}$ is passed directly to the output projection to obtain the final next-token distribution. The selected token is appended to the common prefix and used by all three branches at the next decoding step. Thus, the counterfactual branches function as internal evidence probes rather than independent generators: they maintain their own key–value states but never produce separate token sequences.

Let $C_{1:L}$ denote the cost of the complete original branch, $C_{1:\ell}$ the cost of one counterfactual branch truncated at the evidence-probing layer, and C_{edit} the additional

cost of evidence decomposition, protected projection, final-layer editing, and recomputation of the output logits. The approximate computation at each decoding step is

$$C_{\text{ours}} \approx C_{1:L} + 2C_{1:\ell} + C_{\text{edit}}. \tag{17}$$

Equation (17) consists of one complete original pass, two truncated counterfactual probes, and the additional evidence-editing operations. Because editing is performed at the final transformer layer, the updated representation does not require another propagation through layers $(\ell+1)$:L. The expression describes computational structure rather than directly predicting wall-clock latency, which also depends on branch scheduling, accelerator utilization, and cache management.

In our implementation, the original, visual-counterfactual, and linguistic-counterfactual states are concatenated along the batch dimension and jointly processed through layers 1:ℓ. Each branch maintains its own key–value state within the batched representation. At layer ℓ, the two counterfactual branches terminate, whereas the original branch continues through the remaining layers to obtain $h_t^{o,L}$ and the preliminary next-token distribution.

The response matrices remain on the same GPU as the LVLM and are converted to FP32 before subspace decomposition. Economy-size singular value decomposition is performed using `full_matrices=False`, and only the leading $\widehat{r}_t^k$ valid singular directions are retained for $k \in \{s, v, l\}$. The resulting projectors are applied to the final-layer representation of the original branch. After claim-aware and evidence-preserving editing, the final logits are recomputed directly from $h_t^{*,L}$, without another transformer-layer propagation.

This implementation combines batched evidence probing, early termination of both counterfactual branches, on-device low-rank decomposition, and final-layer editing. It therefore avoids two complete counterfactual forward passes, an additional upper-layer propagation of the edited representation, and CPU–GPU transfers during evidence-subspace construction.

All three branches reuse a single copy of the LVLM parameters. Counterfactual probing therefore does not replicate the model weights. The additional persistent memory is introduced primarily by the two independent key–value caches maintained by the counterfactual branches up to layer ℓ. Let $M_{1:L}^{\text{KV}}$ denote the key–value cache required by the complete original branch and $M_{1:\ell}^{\text{KV}}$ the cache required by one branch truncated at the evidence-probing layer. The total cache requirement is approximated by

$$M_{\text{ours}}^{\text{KV}} \approx M_{1:L}^{\text{KV}} + 2M_{1:\ell}^{\text{KV}}.$$

Because the two counterfactual caches terminate at layer ℓ, their memory cost scales with the probing depth rather than the complete model depth. Under the standard approximation that cache allocation is proportional to the number of transformer layers, the additional cache requirement for LLaVA-1.5-7B with $\ell = 26$ and $L = 32$ is

$$\frac{M_{\text{ours}}^{\text{KV}} - M_{1:L}^{\text{KV}}}{M_{1:L}^{\text{KV}}} \approx \frac{2\ell}{L} = 1.625.$$

Thus, the two truncated counterfactual branches add approximately 162.5% of the greedy-decoding key–value cache, giving a total cache allocation of approximately 262.5% of the greedy baseline. These percentages refer specifically to the key–value cache rather than the complete GPU memory footprint. The LVLM parameters remain a single shared allocation and are not multiplied by the number of branches.

Together, shared model parameters, truncated counterfactual caches, batched probe execution, one-time visual-counterfactual construction, on-device low-rank decomposition, and final-layer editing keep the additional inference cost structurally controlled. The complete procedure remains training-free and requires no parameter updates, auxiliary models, object detectors, or external verification systems. The end-to-end throughput reported in Table 13 provides the empirical counterpart to this analysis by measuring the complete evidence-probing and selective-editing pipeline.

4 Experiments

We evaluate the proposed method along four dimensions: hallucination mitigation, transfer across LVLM backbones, preservation of visual and general multimodal capabilities, and end-to-end inference efficiency. The six benchmarks—POPE, CHAIR, MME, AMBER, HallusionBench, and MM-Vet—span controlled object-existence diagnosis, open-ended image description, fine-grained visual perception, broader visual-consistency assessment, and general multimodal understanding. Beyond headline performance, component ablations, sensitivity analyses, representation-level measurements, and qualitative cases identify which components drive the gains and whether the observed behavior matches the intended evidence-editing mechanism.

4.1 Experimental Setup

We evaluate three representative LVLM families under two complementary backbone regimes. For the controlled POPE diagnostic, we use LLaVA-1.5-7B [1], Qwen2.5-VL-7B [2], and InternVL3-8B [3]. For CHAIR and the fine-grained capability evaluations, we use LLaVA-1.5-7B, Qwen3-VL-8B-Instruct [18], and InternVL3.5-8B [19].

The two regimes serve different but complementary purposes. POPE tests whether the intervention transfers across model families under controlled object-existence questions. CHAIR and the capability benchmarks test faithful open-ended generation and visual-capability preservation on newer generative backbones. LLaVA-1.5-7B is retained as a common reference between the two regimes, while the Qwen and InternVL variants are updated for the more demanding open-ended evaluations. In total, the protocol covers five backbone configurations from three LVLM families. Every method comparison is conducted within the same backbone and benchmark, so the conclusions do not rely on comparing scores across different model generations.

The benchmark suite is structured to distinguish hallucination mitigation from capability preservation. POPE [4] provides a controlled object-existence evaluation under random, popular, and adversarial sampling, for which we report Accuracy and F1. CHAIR [5] measures object hallucination in open-ended descriptions of MS COCO images [20] using CHAIR_i and CHAIR_s, while BLEU examines whether hallucination reduction comes at the cost of caption quality.

The remaining benchmarks test whether this behavior extends beyond controlled object queries and object-centric captioning. We use the Existence, Count, Position, and Color categories of MME [21] to assess fine-grained visual perception; their sum is reported as MME-H. AMBER [22] and HallusionBench [23] provide complementary evaluations of hallucination and visual consistency, while MM-Vet [24] measures general multimodal capability. Taken together, these benchmarks test whether the method suppresses unsupported content without obtaining lower hallucination scores merely by making generation uniformly conservative.

Counterfactual evidence is probed at a comparable relative depth across the evaluated backbones. Specifically, we use the 26th transformer layer of LLaVA-1.5-7B, the 23rd layer of Qwen2.5-VL-7B and InternVL3-8B, and the 29th layer of Qwen3-VL-8B-Instruct and InternVL3.5-8B. All layer indices are one-based. These positions are chosen to obtain intermediate representations in which visual and linguistic evidence has been sufficiently integrated while retaining a clear response to modality-specific perturbations. The resulting evidence subspaces are used to diagnose sample-specific directions, whereas the representation update itself is applied at the final transformer layer.

The visual and linguistic counterfactual retention coefficients are fixed at $\alpha_v = 0.8$ and $\alpha_l = 0.9$, respectively. The slightly higher linguistic retention coefficient applies a milder perturbation to the token-embedding sequence, helping preserve the semantic continuity of the shared decoding prefix. The visual setting applies a stronger perturbation to weaken image-conditioned evidence while retaining the overall input structure. Unless otherwise specified, the subspace rank, claim-triggering threshold, preservation strength, and maximum editing strength are set to $r = 8$, $\tau = 0.5$, $\beta = 0.5$, and $\lambda_{\max} = 1.0$, respectively. The value $\lambda_{\max} = 1.0$ constrains the adaptive editing coefficient to $[0, 1]$. The two retention coefficients are treated as modality-specific probe settings and remain fixed across all backbones and benchmarks. The default values of r, τ, and β are selected from the reported sensitivity analysis.

For each input, one realization of ϵ_v and one realization of ϵ_l are sampled and reused throughout generation; new realizations are drawn for each independent run. Results obtained under our protocol use identical evaluation samples, prompts, and generation lengths and are averaged over three independent runs.

In Tables 1–6, entries marked with † denote results reported in the corresponding cited publications, unmarked entries denote results obtained under our evaluation protocol, and "–" denotes unavailable results. Inference efficiency is measured on a single NVIDIA GeForce RTX 4090 GPU using FP16, an external input batch size of one, and a maximum generation length of 64 tokens. The three branches form an internal batch and are processed jointly up to the evidence-probing layer. Response matrices are cast to FP32, and economy-size SVD is executed on the same GPU using `full_matrices=False`. After ten warm-up samples, TPS is averaged over five runs on 500 images from the MS COCO 2014 validation set. It is computed as the total number of generated tokens divided by elapsed inference time. Timing covers the complete pipeline, including visual encoding, counterfactual probing, subspace construction, claim-aware editing, and autoregressive decoding; only dataset loading and disk I/O are excluded.

Table 1 Results on POPE under random, popular, and adversarial settings. Higher Accuracy (Acc.) and F1 indicate better performance.

Category	Method	LLaVA-1.5-7B		Qwen2.5-VL-7B		InternVL3-8B	
		Acc.↑	F1↑	Acc.↑	F1↑	Acc.↑	F1↑
Random	Greedy	87.8†	87.5†	89.27†	88.14†	89.20†	88.05†
	VCD [6]	88.4†	87.7†	89.60†	88.61†	89.36	88.8
	AGCI [25]	85.57	87.37	91.2	90.9	89.57†	88.53†
	HulluEdit [15]	90.4†	90.5†	91.7	91.3	90.1	89.6
	REVIS [26]	89.37†	88.68†	91.90†	91.43†	90.02	90.38
	Ours	**91.0**	**91.1**	**93.2**	**92.9**	**91.9**	**92.8**
Popular	Greedy	82.5†	83.2†	88.23†	87.14†	88.00†	86.90†
	VCD	83.1†	84.1†	88.27†	87.34†	88.1	87.3
	AGCI	86.87	85.76	89.8	88.9	88.13†	87.16†
	HulluEdit	87.5†	87.6†	90.1	89.2	88.9	87.8
	REVIS	87.13†	86.63†	90.30†	89.91†	89.1	88.2
	Ours	**88.5**	**88.6**	**91.2**	**90.7**	**90.3**	**89.1**
Adversarial	Greedy	77.6†	79.4†	87.10†	86.09†	86.43†	85.43†
	VCD	78.1†	79.6†	87.03†	86.19†	86.9	86.1
	AGCI	82.6	82.7	87.36	86.98	86.57†	85.70†
	HulluEdit	82.5†	83.4†	87.58	87.39	86.9	85.8
	REVIS	83.50†	83.51†	87.80†	87.63†	87.0	86.8
	Ours	**84.6**	**84.9**	**89.2**	**88.9**	**87.9**	**87.0**

4.2 Comparison with State-of-the-Art Methods

Table 1 reports POPE results under random, popular, and adversarial sampling across three LVLM backbones. The proposed method ranks first in both Accuracy and F1 in every backbone–split setting, covering all 18 reported metric comparisons. On the adversarial split, its Accuracy exceeds the strongest competing value reported in the table by 1.1, 1.4, and 0.9 points on LLaVA-1.5-7B, Qwen2.5-VL-7B, and InternVL3-8B, respectively.

Adversarial POPE emphasizes frequent or strongly co-occurring objects that are absent from the image, making these gains particularly relevant to hallucinations driven by linguistic co-occurrence. The first-place ranking across all three backbones demonstrates that the performance gain transfers across the evaluated architectures rather than depending on one model or question distribution. This pattern is consistent with the intended role of sample-specific evidence attribution: a claim is attenuated when its linguistic plausibility is not matched by corresponding visual evidence.

Table 2 evaluates open-ended image description using CHAIR and BLEU. The proposed method achieves the lowest CHAIR_i on all three backbones and reduces both CHAIR_i and CHAIR_s relative to HulluEdit in every case. Specifically, the relative reductions in CHAIR_i are 9.1%, 9.7%, and 14.7%, while those in CHAIR_s are 6.2%, 13.1%, and 7.6% on LLaVA-1.5-7B, Qwen3-VL-8B-Instruct, and InternVL3.5-8B,

Table 2 Results on open-ended image description evaluated by CHAIR and BLEU. Lower CHAIR_i and CHAIR_s indicate fewer hallucinations, while higher BLEU indicates better caption quality.

Method	LLaVA-1.5-7B			Qwen3-VL-8B-Instruct			InternVL3.5-8B		
	$\text{CHAIR}_i \downarrow$	$\text{CHAIR}_s \downarrow$	BLEU ↑	$\text{CHAIR}_i \downarrow$	$\text{CHAIR}_s \downarrow$	BLEU ↑	$\text{CHAIR}_i \downarrow$	$\text{CHAIR}_s \downarrow$	BLEU ↑
Greedy	7.08†	20.40†	15.72†	4.62†	8.40†	9.81†	10.80†	41.40†	**14.0**
VCD [6]	7.28†	20.30†	14.53†	11.70	38.40	9.2	–	–	–
OPERA [9]	6.07†	17.50†	16.02†	4.0	7.2	**10.1**	–	–	–
Nullu [12]	5.30†	15.20†	15.69†	3.5	6.3	9.8	–	–	–
HIME [27]	4.56†	13.80†	15.82†	3.44†	6.00†	8.89†	–	–	–
VAF [11]	6.50	20.40	14.60	9.75†	53.60†	9.3	11.00†	42.20†	13.7
SAVAA [28]	–	–	–	8.38†	46.00†	9.6	9.34†	34.40†	12.9
TITA [29]	5.60†	16.30†	**16.65**	12.20†	20.30†	9.4	8.50†	**22.30**†	12.7
HulluEdit [15]	4.18†	13.00†	15.49†	3.1	6.1	9.7	7.5	28.38	12.3
Ours	**3.80**	**12.20**	15.30	**2.8**	**5.3**	9.7	**6.4**	26.23	12.4

respectively. The largest CHAIR_i reduction is obtained on InternVL3.5-8B, showing that the gain extends to a recent and architecturally distinct backbone.

The reduction in hallucination is achieved without a corresponding loss in caption quality. Relative to HulluEdit, BLEU changes by only -0.19, 0.00, and $+0.10$ points on the three backbones, respectively. This combination of consistently lower hallucination scores and essentially unchanged BLEU shows that the method improves visual faithfulness without making open-ended descriptions less informative. It therefore achieves the central objective of selective evidence editing: suppressing unsupported claims while preserving useful descriptive content.

4.3 Generalization and Capability Preservation

Having established consistent hallucination reduction on POPE and CHAIR, we next examine whether these gains are accompanied by strong visual and multimodal capability. Tables 3–6 evaluate this joint objective across fine-grained visual perception, information coverage, visual consistency, and general multimodal capability. Together, these benchmarks determine whether the proposed method improves visual faithfulness while retaining the information required for accurate and informative generation.

Table 3 provides the first test through the Existence, Count, Position, and Color categories of MME. Summing these four categories yields MME-H scores of 590.00, 775.00, and 771.67 on LLaVA-1.5-7B, Qwen3-VL-8B-Instruct, and InternVL3.5-8B, respectively, ranking the proposed method first on every backbone. Compared with the strongest aggregate baseline for each backbone, these scores represent gains of 3.33, 15.00, and 8.33 points, respectively.

The category-level results reinforce this aggregate advantage. On Qwen3-VL-8B-Instruct and InternVL3.5-8B, the proposed method matches or exceeds the best baseline in every evaluated category. On LLaVA-1.5-7B, it achieves the highest aggregate MME-H score while attaining the best or tied-best results in Existence, Position, and Color.

Together, these results demonstrate that the proposed method combines consistent hallucination mitigation with preserved and often improved fine-grained visual

Table 3 Results on fine-grained visual capability evaluation using MME. Higher values indicate better performance.

Backbone	Method	Existence↑	Count↑	Position↑	Color↑
LLaVA-1.5-7B	Vanilla	181.67†	118.33†	104.44†	152.78†
	Nullu [12]	190.00†	121.11†	105.56†	156.67†
	DeCo [30]	175.00†	**128.33**†	98.33†	125.00†
	HulluEdit [15]	**195.00**†	105.00†	126.67†	160.00†
	Ours	**195.00**	105.00	**128.33**	**161.67**
Qwen3-VL-8B-Instruct	Vanilla	195.00†	175.00†	168.33†	193.33†
	VCD [6]	195.00†	180.00†	168.33†	193.33†
	OPERA [9]	195.00†	180.00†	168.33†	**200.00**†
	Kestrel [31]	**200.00**†	186.67†	180.00†	193.33†
	Ours	**200.00**	**190.00**	**185.00**	**200.00**
InternVL3.5-8B	Vanilla	**200.00**†	175.00†	175.00†	193.33†
	VCD	**200.00**†	175.00†	175.00†	193.33†
	OPERA	195.00†	173.33†	175.00†	195.00†
	Kestrel	**200.00**†	186.67†	**181.67**†	195.00†
	Ours	**200.00**	**190.00**	**181.67**	**200.00**

Table 4 Results on the AMBER hallucination benchmark.

Method	CHAIR ↓	Coverage ↑	Hallucination ↓	Cognition ↓
Vanilla	8.2†	**48.9**†	34.3†	4.0†
DeCo [30]	6.6†	47.5†	28.1†	2.8†
HulluEdit [15]	5.3	47.0	22.4	2.1
Ours	**4.7**	48.1	**20.6**	**1.8**

perception. This performance pattern is consistent with the evidence-preserving objective: unsupported evidence is selectively attenuated without broadly suppressing the visual signals required for accurate perception.

Table 4 extends the evaluation to AMBER. The proposed method achieves the lowest CHAIR, Hallucination, and Cognition scores among all competing methods. Relative to HulluEdit, it reduces these three metrics by 11.3%, 8.0%, and 14.3%, respectively.

These reductions are achieved without sacrificing information coverage. Coverage increases by 1.1 points over HulluEdit, from 47.0 to 48.1, and retains 98.4% of the coverage achieved by vanilla decoding. This combination of consistently lower hallucination scores and near-vanilla coverage demonstrates that the method improves reliability while preserving the breadth of visually relevant content. The result is consistent with the proposed selective evidence-editing design, which attenuates unsupported evidence while keeping useful image-grounded information available for generation.

Table 5 Results on HallusionBench. Higher values indicate better performance.

Method	Overall↑	qAcc↑	fAcc↑	aAcc↑
Vanilla	27.6	13.6	20.5	48.8
QLoRA [32]	25.2	13.2	16.2	46.2
HulluEdit [15]	30.1	16.9	21.4	52.1
Ours	**31.4**	**17.6**	**22.3**	**53.5**

Table 6 Results on MM-Vet for general multimodal capability evaluation.

Model	Method	MM-Vet Total↑
LLaVA-1.5-7B	Vanilla	23.6†
LLaVA-1.5-7B	DeCo [30]	27.9†
LLaVA-1.5-7B	HulluEdit [15]	28.5
LLaVA-1.5-7B	Ours	**29.3**

Table 5 examines whether the advantage of the proposed method extends beyond object-centric evaluation to the broader visual-consistency scenarios represented by HallusionBench. The proposed method ranks first on all four reported metrics. Compared with HulluEdit, the strongest competing method in this table, it improves Overall, qAcc, fAcc, and aAcc by 1.3, 0.7, 0.9, and 1.4 points, respectively. The advantage therefore spans every evaluation component rather than being driven by a single metric.

Together with the POPE and CHAIR results, this across-metric advantage demonstrates that the proposed method generalizes from controlled object-existence diagnosis and open-ended image description to broader visual-consistency evaluation. These results provide strong evidence that sample-specific evidence attribution remains effective across distinct hallucination settings rather than being tied to a single benchmark or evaluation regime. We further evaluate general multimodal capability on MM-Vet in Table 6. The proposed method achieves the highest score of 29.3, exceeding HulluEdit by 0.8 points and vanilla decoding by 5.7 points, which corresponds to a 24.2% relative improvement over vanilla decoding. Combined with the preceding hallucination results, this outcome confirms that the reduction in unsupported content is not achieved at the expense of general multimodal performance. Instead, the method simultaneously delivers stronger hallucination mitigation and the best MM-Vet result among the compared approaches, supporting the effectiveness of its evidence-preserving selective-editing strategy. Taken together, Tables 3–6 establish that hallucination mitigation and capability preservation are not competing outcomes for the proposed method. It achieves the highest aggregate MME-H score on all three backbones, the lowest values on all three hallucination-oriented AMBER metrics while retaining 98.4% of vanilla coverage, the highest score on every HallusionBench metric,

and the best MM-Vet result. This consistency across fine-grained perception, information coverage, visual consistency, and general multimodal capability demonstrates that the framework suppresses unsupported content without making generation uniformly conservative. The method therefore improves visual reliability while preserving, and in several settings improving, the evaluated capabilities of the underlying LVLM.

4.4 Ablation and Sensitivity Analysis

Table 7 isolates the contribution of each component on LLaVA-1.5-7B. The CHAIR_i and CHAIR_s metrics quantify hallucination, whereas MME-H measures fine-grained visual capability. Evaluating these objectives together reveals not only which components drive hallucination reduction, but also whether that reduction is obtained at the cost of visual perception.

Dual-counterfactual probing is the dominant contributor to hallucination mitigation. Removing both probes produces the worst CHAIR results among all ablated variants: CHAIR_i increases from 3.80 to 4.86 and CHAIR_s from 12.20 to 14.42, corresponding to relative degradations of 27.9% and 18.2%, respectively. MME-H simultaneously decreases by 16.67 points, from 590.00 to 573.33. The single-probe variants recover part of the performance; although L-CF only outperforms V-CF only on all three metrics, neither matches the full model. These results establish that the two counterfactual probes provide complementary evidence. The visual probe measures sensitivity to image evidence, while the linguistic probe measures sensitivity to the aligned textual context; their paired responses enable sample-specific attribution of the two evidence sources.

Shared-exclusive decomposition determines the evidence structure presented to the selective editor. Removing this decomposition degrades all three metrics: CHAIR_i increases from 3.80 to 4.46, CHAIR_s increases from 12.20 to 13.72, and MME-H decreases by 13.33 points, from 590.00 to 576.67. Retaining only the shared component produces the weakest subspace-decomposition variant, with scores of 4.63, 14.05, and 575.00. Using only the exclusive components performs better, but still remains below the full model on every metric. This ordered pattern demonstrates that neither cross-modal agreement nor modality-exclusive residuals alone are sufficient; their joint decomposition is necessary to construct a complete sample-specific evidence profile.

Claim-aware and adaptive editing determine when the identified evidence should be modified and how strongly the intervention should be applied. Removing claim-aware editing produces the largest deterioration within the editing-mechanism group, increasing CHAIR_i and CHAIR_s to 4.31 and 13.40 while reducing MME-H to 580.00. Token-level editing is the strongest simplified alternative, but it still reaches only 4.12, 12.98, and 583.33, respectively. A fixed editing strength likewise underperforms the complete method on all three metrics. The consistent gap across these variants demonstrates that both claim selectivity and evidence-dependent editing strength are required to achieve the best balance between hallucination mitigation and capability preservation.

Evidence preservation is the decisive component for capability retention. Removing the preservation constraint causes the largest MME-H loss among all ablated variants, decreasing the score by 21.67 points from 590.00 to 568.33. Hallucination performance

Table 7 Ablation study of different components. Lower CHAIR_i and CHAIR_s indicate fewer hallucinations, while higher MME-H indicates better preservation of fine-grained visual capabilities.

Category	Variant	$\text{CHAIR}_i \downarrow$	$\text{CHAIR}_s \downarrow$	MME-H↑
Full Model	Full Model	**3.80**	**12.20**	**590.00**
Counterfactual Construction	V-CF only	4.28	13.35	583.33
	L-CF only	4.17	13.08	585.00
	w/o Dual-CF	4.86	14.42	573.33
Subspace Decomposition	w/o Shared-Exclusive Decomposition	4.46	13.72	576.67
	Shared Component only	4.63	14.05	575.00
	Exclusive Components only	4.34	13.51	578.33
Editing Mechanism	w/o Claim-Aware Editing	4.31	13.40	580.00
	Token-Level Editing	4.12	12.98	583.33
	Fixed Editing Strength	4.24	13.26	581.67
Preservation Mechanism	w/o Preservation Constraint	4.39	13.61	568.33
	Fixed Preservation Strength	4.09	12.87	585.00

also deteriorates, with CHAIR_i and CHAIR_s increasing from 3.80 and 12.20 to 4.39 and 13.61, respectively. A fixed preservation strength partially recovers the three metrics to 4.09, 12.87, and 585.00, but the complete preservation mechanism improves all of them further. Evidence preservation therefore does more than retain visual capability: it enables hallucination-oriented editing to remain selective rather than disturbing reliable visual support.

Taken together, the component ablations establish a clear division of responsibility. Dual-counterfactual probing diagnoses modality dependence, shared-exclusive decomposition organizes the resulting evidence, claim-aware adaptive editing determines when and how strongly to intervene, and evidence preservation protects reliable visual information. No reduced variant matches the complete method on all three metrics, demonstrating that these components operate as a coherent evidence-editing pipeline. Having established their individual contributions, we next examine sensitivity to the evidence-probing layer, subspace rank r, claim threshold τ, and preservation coefficient β. All numerical sensitivity experiments are conducted on LLaVA-1.5-7B.

The perturbation coefficients α_v and α_l control the proportion of the original input retained by the visual and linguistic counterfactual branches. According to Eqs. (4) and (5), as either coefficient approaches one, the corresponding counterfactual becomes increasingly similar to the original input and produces a weaker response. Conversely, reducing the coefficient strengthens the counterfactual perturbation; an excessively small value may disrupt the underlying semantic structure and cause the measured response to reflect arbitrary corruption rather than controlled evidence weakening.

We therefore adopt the moderate asymmetric setting $\alpha_v = 0.8$ and $\alpha_l = 0.9$. The lower visual coefficient produces a sufficiently distinct visual probe, whereas the more conservative linguistic coefficient better preserves the semantic structure of the aligned prefix and its autoregressive decoding context. Once specified, both coefficients remain fixed across all evaluated backbones and benchmarks. The reported comparisons therefore do not rely on benchmark-specific perturbation tuning.

With the editing layer fixed at the final transformer layer, Table 8 evaluates the layer from which the counterfactual evidence responses are extracted. Among the

Table 8 Effect of the evidence-probing layer on hallucination mitigation. Lower CHAIR_i and CHAIR_s indicate better performance.

Evidence-Probing Layer	Editing Layer	$\text{CHAIR}_i \downarrow$	$\text{CHAIR}_s \downarrow$
Layer 20	Last	4.48	14.73
Layer 22	Last	4.21	13.65
Layer 24	Last	3.96	12.72
Layer 26	Last	**3.80**	**12.20**
Layer 28	Last	3.91	12.55
Layer 30	Last	4.12	13.06
Last	Last	4.42	14.18

Table 9 Sensitivity analysis of the rank parameter r used for evidence-subspace construction.

r	$\text{CHAIR}_i \downarrow$	$\text{CHAIR}_s \downarrow$	MME-H↑
4	4.02	12.91	583.33
8	**3.80**	**12.20**	**590.00**
12	3.87	12.42	588.33
16	4.05	12.88	585.00

evaluated settings, Layer 26 provides the best evidence-probing position, where both CHAIR metrics attain their minimum values. Moving the probe from Layer 20 to Layer 26 reduces CHAIR_i from 4.48 to 3.80 and CHAIR_s from 14.73 to 12.20, corresponding to relative improvements of 15.2% and 17.2%, respectively. Both metrics improve consistently through Layer 26 and then deteriorate at deeper positions; compared with final-layer probing, Layer 26 lowers both scores by approximately 14.0%. This depth-dependent pattern is consistent with the progressive organization of multimodal evidence: earlier layers may not yet integrate the two modalities sufficiently, whereas output-adjacent representations are more specialized for next-token prediction and provide less separable evidence responses. We therefore use Layer 26 as the default evidence-probing position for LLaVA-1.5-7B. Table 9 evaluates the sensitivity to the subspace rank r. Among the evaluated ranks, $r = 8$ provides the best joint result across all three metrics. Compared with $r = 4$, this setting reduces both CHAIR_i and CHAIR_s by approximately 5.5% and increases MME-H from 583.33 to 590.00, a gain of 6.67 points. Increasing the rank beyond 8 partially reverses these gains; at $r = 16$, the two CHAIR scores rise to 4.05 and 12.88, while MME-H decreases to 585.00. Thus, simply enlarging the evidence subspace does not improve the intervention. This non-monotonic pattern is consistent with $r = 8$ retaining the dominant sample-specific evidence directions, whereas additional lower-energy directions may reduce the selectivity of the resulting projection. We therefore adopt $r = 8$ as the default rank parameter. Table 10 characterizes the trade-off controlled by the claim-triggering threshold τ. A lower threshold activates the claim gate more broadly: at $\tau = 0.30$, the method obtains the lowest CHAIR_i and CHAIR_s scores of 3.68 and

Table 10 Sensitivity analysis of the claim-triggering threshold τ.

τ	$\text{CHAIR}_i \downarrow$	$\text{CHAIR}_s \downarrow$	MME-H↑
0.30	**3.68**	**11.90**	575.00
0.40	3.74	12.03	584.00
0.50	3.80	12.20	590.00
0.60	4.01	12.76	591.67
0.70	4.27	13.31	**593.33**

Table 11 Sensitivity analysis of the preservation strength β.

β	$\text{CHAIR}_i \downarrow$	$\text{CHAIR}_s \downarrow$	MME-H↑
0	4.39	13.61	568.33
0.25	4.06	12.77	581.67
0.50	**3.80**	**12.20**	590.00
0.75	3.92	12.51	592.00
1.00	4.11	12.96	**594.00**

11.90, respectively, but MME-H decreases to 575.00. Raising the threshold progressively restricts intervention and restores visual capability, reaching an MME-H score of 593.33 at $\tau = 0.70$, but also increases the two CHAIR scores to 4.27 and 13.31. The setting $\tau = 0.50$ provides the most favorable operating balance. Relative to $\tau = 0.30$, it recovers 15.00 MME-H points while increasing CHAIR_i and CHAIR_s by only 0.12 and 0.30. In contrast, increasing τ from 0.50 to 0.70 yields only 3.33 additional MME-H points but raises the two CHAIR scores by 0.47 and 1.11. We therefore adopt $\tau = 0.50$ as the default claim-triggering threshold.

Table 11 shows that evidence preservation is not simply a constraint that weakens intervention. Increasing β from 0 to 0.5 improves all three metrics: CHAIR_i and CHAIR_s decrease by 13.4% and 10.4%, respectively, while MME-H increases by 21.67 points from 568.33 to 590.00. This joint improvement indicates that filtering directions that overlap with protected evidence makes the intervention more selective rather than merely reducing its magnitude. Increasing β beyond 0.5 further improves MME-H, with full preservation reaching the maximum score of 594.00 at $\beta = 1$, but the two CHAIR scores deteriorate to 4.11 and 12.96. The default setting $\beta = 0.5$ retains 99.3% of the maximum MME-H score while achieving the lowest CHAIR_i and CHAIR_s. It therefore provides the strongest balance between hallucination mitigation and visual-capability preservation. Taken together, Tables 8–11 define a stable and interpretable operating region for the proposed method. Layer 26 and $r = 8$ provide the best results within the evaluated ranges for evidence probing and subspace construction, whereas $\tau = 0.5$ and $\beta = 0.5$ provide the strongest balance between hallucination mitigation and visual-capability preservation. Neighboring settings produce gradual rather than abrupt changes in the evaluation metrics, showing that the method does not depend

on a narrowly tuned configuration. Accordingly, the default configuration uses $r = 8$, $\tau = 0.5$, and $\beta = 0.5$, with $\ell = 26$ used as the evidence-probing layer for LLaVA-1.5-7B.

4.5 Mechanism and Efficiency Analysis

To quantify how selective editing affects each evidence component, we first define the projected components of the original and edited final-layer representations. For $k \in \{s, v, l\}$,

$$c_t^{k,o} = P_t^k h_t^{o,L}, \qquad c_t^{k,*} = P_t^k h_t^{*,L}.$$

The residual component is defined as the orthogonal complement of the union of the shared, visual-exclusive, and linguistic-exclusive evidence subspaces:

$$U_t^e = \mathrm{Orth}\left(\left[U_{t,\widehat{r}_t^s}^s, U_{t,\widehat{r}_t^v}^v, U_{t,\widehat{r}_t^l}^l\right]\right), \qquad P_t^r = I_d - U_t^e (U_t^e)^\top.$$

Accordingly,

$$c_t^{r,o} = P_t^r h_t^{o,L}, \qquad c_t^{r,*} = P_t^r h_t^{*,L}.$$

Here, s, v, l, and r denote shared, visual-exclusive, linguistic-exclusive, and residual evidence, respectively. For the set $\mathcal{T}$ of claim-triggered decoding steps, we compute the mean relative component change as

$$\Delta_k = \frac{100\%}{|\mathcal{T}|} \sum_{t \in \mathcal{T}} \frac{\left\| c_t^{k,*} - c_t^{k,o} \right\|_2}{\left\| c_t^{k,o} \right\|_2 + \varepsilon}.$$

The evidence projectors are estimated at the probing layer ℓ and applied to the original and edited final-layer representations. Thus, Δ_k measures how strongly final-layer editing changes the evidence associated with each sample-specific subspace. All reported values are averaged over the evaluated samples and their claim-triggered decoding steps.

Table 12 provides representation-level evidence that the proposed update is concentrated on linguistic-exclusive evidence rather than broadly perturbing the hidden state. With the complete model, Δ_l remains substantial at 31.6%, while Δ_s, Δ_v, and Δ_r are limited to 2.1%, 2.8%, and 9.8%, respectively. Relative to the variant without preservation, the adaptive preservation strategy reduces the changes in shared and visual-exclusive evidence by 76.4% and 77.4%, while retaining 92.4% of the intervention magnitude on linguistic-exclusive evidence. This asymmetric change pattern directly matches the intended behavior of evidence-preserving editing: the update remains strong in the target subspace but is redirected away from protected evidence.

The increased selectivity is also reflected in downstream performance. Removing preservation raises CHAIR_i and CHAIR_s from 3.80 and 12.20 to 4.39 and 13.61, while reducing MME-H from 590.00 to 568.33. Fixed preservation produces intermediate representation changes and intermediate benchmark performance, whereas the complete model achieves the best CHAIR and MME-H results. Crucially, this improvement is obtained without maximizing Δ_l. The results therefore demonstrate that adaptive preservation redirects the edit toward unsupported linguistic evidence rather

Table 12 Validation of the evidence-subspace mechanism and preservation strategy. Δ_s, Δ_v, Δ_l, and Δ_r denote the relative changes in the shared, visual-exclusive, linguistic-exclusive, and residual evidence components, respectively. Lower Δ_s, Δ_v, and Δ_r indicate less disturbance to non-target evidence, while Δ_l measures the intervention magnitude on linguistic-exclusive evidence and has no monotonic optimum. Lower CHAIR scores and higher MME-H indicate better hallucination mitigation and visual-capability preservation, respectively.

Variant	$\Delta_s \downarrow$	$\Delta_v \downarrow$	Δ_l	$\Delta_r \downarrow$	$\text{CHAIR}_i \downarrow$	$\text{CHAIR}_s \downarrow$	MME-H $\uparrow$
w/o Preservation	8.9%	12.4%	34.2%	12.1%	4.39	13.61	568.33
Fixed Preservation	4.6%	6.1%	32.8%	10.7%	4.09	12.87	585.00
Full Model	**2.1%**	**2.8%**	31.6%	**9.8%**	**3.80**	**12.20**	**590.00**

Table 13 End-to-end inference efficiency measured by tokens per second (TPS) and throughput relative to greedy decoding. All measurements include the complete inference pipeline; higher values indicate faster inference.

Method	TPS (tokens/s) $\uparrow$	Relative throughput (%) $\uparrow$
Greedy	**31.25**	**100.0**
Beam Search	29.91	95.7
DoLa [33]	23.61	75.6
HulluEdit [15]	20.02	64.1
VCD [6]	17.58	56.3
OPERA [9]	4.36	14.0
HALC [7]	2.15	6.9
Ours	23.59	75.5

than uniformly weakening the intervention. Table 13 demonstrates that the complete evidence-editing pipeline retains practical end-to-end speed on a single NVIDIA GeForce RTX 4090 GPU. The proposed method reaches 23.59 tokens per second, preserving 75.5% of greedy-decoding throughput while performing dual-counterfactual probing, sample-specific evidence decomposition, and selective representation editing. Its throughput is effectively identical to that of DoLa, differing by only 0.02 tokens per second. It is also 17.8% and 34.2% faster than HulluEdit and VCD, respectively, and achieves 5.4× and 11.0× the throughput of OPERA and HALC.

The favorable runtime profile follows from the partial, on-device implementation. The three branches are processed as an internal batch only up to the evidence-probing layer; the two counterfactual branches then terminate, while the visual counterfactual is constructed once and reused throughout generation. The response matrices and their economy-size SVD computations also remain on the GPU. Importantly, the reported throughput includes counterfactual probing, subspace construction, representation editing, and autoregressive decoding, rather than measuring an isolated editing operation. The results therefore establish that sample-specific evidence attribution can be integrated into LVLM decoding while retaining three quarters of greedy-decoding throughput and remaining faster than several established inference-time mitigation methods.

4.6 Qualitative Analysis

Building on the consistent quantitative gains across six benchmarks, Fig. 2 visualizes how the proposed method achieves hallucination reduction without sacrificing descriptive completeness across four representative error types: object existence, quantity, visual attributes, and spatial relations. Greedy decoding produces a distinct unsupported claim in each example, including a nonexistent backpack, an incorrect horse count, inaccurate attributes of the cat, and a reversed spatial relation between the cyclist and the divider. HulluEdit removes the erroneous content but, in several cases, also omits visually supported details, resulting in less informative descriptions.

The outputs produced by our method satisfy both objectives in all four examples: the unsupported claims are removed while grounded details are retained, including the rider's clothing and mountainous background, the two horses and the church, the cat's orange-and-white long-haired appearance, and the cyclist's correct position relative to the divider. The key distinction is therefore not merely whether a hallucination is suppressed, but whether reliable visual evidence survives the correction. These examples make the aggregate benchmark results concrete: the proposed method consistently suppresses unsupported linguistic evidence while preserving informative visual details. Together, the quantitative and qualitative results demonstrate that the improvement arises from selective evidence correction rather than globally conservative generation.

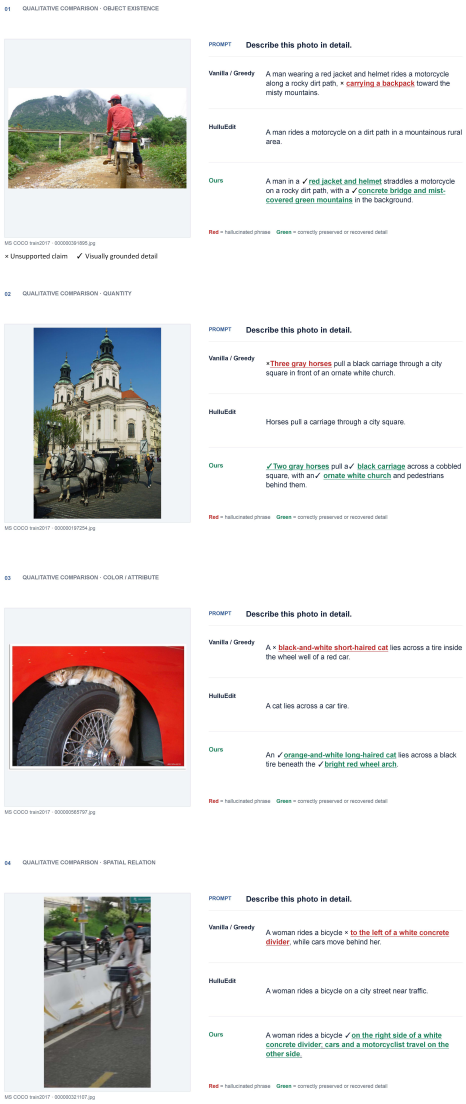


Fig. 2 Qualitative comparisons among Greedy decoding, HulluEdit, and the proposed method on four images from the MS COCO 2017 training split [20]. The cases cover object existence, quantity, visual attributes, and spatial relations. A cross together with red underlining marks unsupported content, whereas a check mark together with green underlining indicates visually grounded details preserved or recovered by our method. Image identifiers are shown below the corresponding images

4.7 Discussion and Extensions

The results across three LVLM families and six complementary benchmarks establish the broad applicability of the proposed evidence-attribution principle. Its central advantage is to treat hallucination mitigation as evidence selection during decoding

rather than as uniform output suppression. By separating shared, visual-exclusive, and linguistic-exclusive evidence for each sample, the framework supports a common intervention rule across object existence, attributes, quantities, spatial relations, observable actions, and open-ended visual descriptions. The consistent gains on POPE, CHAIR, MME, AMBER, HallusionBench, and MM-Vet show that this principle transfers across model architectures and evaluation settings while preserving visual and general multimodal capability.

The framework is also modular by construction. The fixed construction of V_{claim} provides a deterministic, benchmark-independent, and computationally lightweight gate that decides when editing is activated, while dual-counterfactual decomposition independently determines which evidence should be attenuated or protected. The current implementation focuses on visual claims that can be identified at a single decoding step. Because claim detection and evidence attribution are modular, the framework can be extended to compositional claims spanning multiple tokens without modifying the underlying evidence decomposition or representation update. Importantly, however, claim detection and evidence attribution remain separate components of the framework.

This separation provides a natural route for extending claim coverage. The fixed vocabulary can be replaced or augmented by context-aware vocabulary construction, phrase-level visual-claim detection, or soft semantic matching without changing the underlying evidence decomposition and final-layer representation update. Similarly, counterfactual responses can be aggregated over phrase- or claim-level spans before shared–exclusive decomposition, allowing the method to attribute evidence distributed across multiple tokens. Such extensions could broaden selective editing toward compositional, causal, and multi-step visual reasoning while retaining the framework's modular, sample-specific, evidence-preserving, and training-free design. Their effectiveness in these broader reasoning settings requires dedicated future evaluation.

5 Conclusion

This work establishes sample-specific evidence attribution as an effective principle for training-free hallucination mitigation in LVLMs. Instead of relying solely on output-distribution contrast or a globally fixed representation edit, the proposed framework uses paired visual and linguistic counterfactuals to diagnose the evidence supporting each visually verifiable claim. It then decomposes the resulting hidden-state responses into shared, visual-exclusive, and linguistic-exclusive components, enabling claim-aware intervention to suppress unsupported linguistic evidence while protecting visually grounded information.

Across three LVLM families and six benchmarks, the proposed method consistently improves visual faithfulness without sacrificing descriptive quality or general multimodal capability. It ranks first in every POPE backbone–split setting and reduces CHAIR_i by 9.1%–14.7% relative to HulluEdit. Complementary capability evaluations, component ablations, and representation-level analysis further show that the gains

arise from selective attenuation of linguistic-exclusive evidence while shared and visual-exclusive information is preserved. The complete pipeline remains practical, reaching 23.59 tokens per second on a single RTX 4090 GPU.

Taken together, these results show that reliable LVLM decoding does not require uniformly conservative generation. Separating evidence diagnosis from selective intervention enables the model to remove unsupported claims while retaining useful visual content. The proposed framework therefore combines consistent hallucination reduction, capability preservation, and practical inference efficiency in a training-free design, providing a modular foundation for more reliable visually grounded generation.

Statements and Declarations

Funding. No funding was received for conducting this study.

Competing Interests. The authors have no relevant financial or non-financial interests to disclose.

Author Contributions. Zhiyong Jia: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Data curation, Visualization, and Writing–original draft. Ziqiang Ma: Conceptualization, Methodology, Supervision, Project administration, and Writing–review and editing. Jie Guo: Validation, Formal analysis, and Writing–review and editing. All authors read and approved the final manuscript.

Data Availability. No new datasets were generated or collected for this study. The experiments use the publicly available POPE [4], MS COCO [20], MME [21], AMBER [22], HallusionBench [23], and MM-Vet [24] benchmarks. These datasets can be accessed through the official repositories provided by their corresponding publications.

Ethics Approval. Not applicable. This study involved no human participants or animals and used only publicly available benchmark datasets.

Consent to Participate. Not applicable.

Consent for Publication. Not applicable.

References

[1] Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 26296–26306 (2024). https://doi.org/10.1109/CVPR52733.2024.02484

[2] Bai, S., et al.: Qwen2.5-VL Technical Report (2025). https://arxiv.org/abs/2502.13923

[3] Zhu, J., et al.: InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models (2025). https://arxiv.org/abs/2504.10479

[4] Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.-R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, pp. 292–305 (2023). https://doi.org/10.18653/v1/2023.emnlp-main.20

[5] Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K.: Object hallucination in image captioning. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 4035–4045 (2018). https://doi.org/10.18653/v1/D18-1437

[6] Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., Bing, L.: 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). https://doi.org/10.1109/CVPR52733.2024.01316

[7] Chen, Z., Zhao, Z., Luo, H., Yao, H., Li, B., Zhou, J.: HALC: Object hallucination reduction via adaptive focal-contrast decoding. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 7824–7846 (2024). https://proceedings.mlr.press/v235/chen24bi.html

[8] Suo, W., Zhang, L., Sun, M., Wu, L.Y., Wang, P., Zhang, Y.: Octopus: Alleviating hallucination via dynamic contrastive decoding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 29904–29914 (2025). https://doi.org/10.1109/CVPR52734.2025.02783

[9] Huang, Q., Dong, X., Zhang, P., Wang, B., He, C., Wang, J., Lin, D., Zhang, W., Yu, N.: 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). https://doi.org/10.1109/CVPR52733.2024.01274

[10] Li, J., Zhang, J., Jie, Z., Ma, L., Li, M., Luo, X., Li, G.: Cross-modal attention calibration for LVLM hallucination mitigation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 40186–40196 (2026)

[11] Yin, H., Si, G., Wang, Z.: 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). https://doi.org/10.1109/CVPR52734.2025.01363

[12] Yang, L., Zheng, Z., Chen, B., Zhao, Z., Lin, C., Shen, C.: Nullu: Mitigating object hallucinations in large vision-language models via HalluSpace projection. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 14635–14645 (2025). https://doi.org/10.1109/CVPR52734.2025.

01364

[13] Li, Z., Shi, H., Gao, Y., Liu, D., Wang, Z., Chen, Y., Liu, T., Zhao, L., Wang, H., Metaxas, D.N.: The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In: Proceedings of the 42nd International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 267, pp. 35799–35819 (2025). https://proceedings.mlr.press/v267/li25ca.html

[14] Li, X., Lin, L., Jiao, H., Yao, L., Wong, T.-T., Wu, H.: 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, pp. 37552–37568 (2026). https://doi.org/10.18653/v1/2026.acl-long.1742

[15] Lin, Y., Fang, Q., Li, Y., Sun, J., Gao, J., Sang, J.: 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)

[16] Zhou, Y., Yang, M., Zhang, J., Li, J., Feng, S.: ALD^2: Adaptive layer-wise denoising decoding for hallucinations mitigation in large vision-language models. Information Processing & Management **63**(7), 104869 (2026) https://doi.org/10.1016/j.ipm.2026.104869

[17] Li, X., Wang, B., Ju, X., Hu, Y., Wang, W.: Training-free counterfactual hallucination mitigation method for large vision-language models. Multimedia Systems **32**, 525 (2026) https://doi.org/10.1007/s00530-026-02577-3

[18] Bai, S., et al.: Qwen3-VL Technical Report (2025). https://arxiv.org/abs/2511.21631

[19] Wang, W., et al.: InternVL3.5: Advancing Open-Source Multimodal Models in Versatility, Reasoning, and Efficiency (2025). https://arxiv.org/abs/2508.18265

[20] Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft COCO: Common objects in context. In: Computer Vision – ECCV 2014. Lecture Notes in Computer Science, vol. 8693, pp. 740–755. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_48

[21] Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., Wu, Y., Ji, R., Shan, C., He, R.: MME: A comprehensive evaluation benchmark for multimodal large language models. In: Advances in Neural Information Processing Systems, vol. 38 (2025). https://doi.org/10.52202/085713-4899

[22] Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., Wang, J., Xu, H., Yan,

M., Zhang, J., Sang, J.: AMBER: An LLM-Free Multi-Dimensional Benchmark for MLLMs Hallucination Evaluation (2023). https://arxiv.org/abs/2311.07397

[23] Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., Manocha, D., Zhou, T.: 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). https://doi.org/10.1109/CVPR52733.2024.01363

[24] Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: MM-Vet: Evaluating large multimodal models for integrated capabilities. In: Proceedings of the 41st International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 235, pp. 57730–57754 (2024). https://proceedings.mlr.press/v235/yu24o.html

[25] Zhu, Y., Bai, X., Chen, K., Xiang, Y., Pan, Y., Hou, Y., Guan, W., Yu, J., Zhang, M.: Beyond the Vision Encoder: Identifying and Mitigating Spatial Bias in Large Vision-Language Models (2025). https://arxiv.org/abs/2509.21984

[26] Wu, J., Shi, W., Shen, H., Qi, P., Tang, K., Huang, Z., Wang, B., Yang, Z.: REVIS: Sparse latent steering to mitigate object hallucination in large vision-language models. In: Proceedings of the 43rd International Conference on Machine Learning (2026). https://openreview.net/forum?id=92tssj0P1S

[27] Akl, A., Khamis, A., Cheraghian, A., Wang, Z., Khalifa, S., Wang, K.: HIME: Mitigating Object Hallucinations in LVLMs via Hallucination Insensitivity Model Editing (2026). https://arxiv.org/abs/2602.18711

[28] Zhang, J., Liu, F., Du, C., Pang, T.: SAVAA: Mitigating Hallucinations in LVLMs via Step-wise Adaptive Visual Attention Amplification (2026). https://arxiv.org/abs/2602.13600

[29] Chen, K., Zheng, J., Zhang, J., Lin, M., Pan, X., Hu, J., Lou, J., Feng, Z., Song, M.: Token-level inference-time alignment for vision-language models. In: Findings of the Association for Computational Linguistics: ACL 2026, pp. 25012–25029 (2026). https://doi.org/10.18653/v1/2026.findings-acl.1253

[30] Wang, C., Chen, X., Zhang, N., Tian, B., Xu, H., Deng, S., Chen, H.: MLLM can see? dynamic correction decoding for hallucination mitigation. In: The Thirteenth International Conference on Learning Representations (2025). https://openreview.net/forum?id=4z3IguA4Zg

[31] Mao, J., Chen, H., Tu, H., Wang, Y., Zhang, L., Zheng, Z., Yao, H., Wang, Z., Xie, C., Zhou, Y.: Kestrel: Grounding Self-Refinement for LVLM Hallucination Mitigation (2026). https://arxiv.org/abs/2603.16664

[32] Dettmers, T., Pagnoni, A., Holtzman, A., Zettlemoyer, L.: QLoRA: Efficient finetuning of quantized LLMs. In: Advances in Neural Information Processing Systems, vol. 36, pp. 10088–10115 (2023). https://doi.org/10.52202/075280-0441

[33] Chuang, Y.-S., Xie, Y., Luo, H., Kim, Y., Glass, J.R., He, P.: DoLa: Decoding by contrasting layers improves factuality in large language models. In: The Twelfth International Conference on Learning Representations (2024). https://openreview.net/forum?id=Th6NyL07na

Author Biographies



Zhiyong Jia is a master's student at the School of Information Engineering, Ningxia University, Yinchuan, China. His research focuses on hallucination analysis and mitigation in multimodal large language models, particularly on detecting and suppressing visually unsupported content. He is also interested in improving factual consistency, visual grounding, and the reliability of multimodal understanding and generation.

E-mail: 15969742062@163.com



Ziqiang Ma received the B.E. degree in electronic information science and technology from Tsinghua University, Beijing, China, in 2012 and the Ph.D. degree in information security from the Institute of Information Engineering, Chinese Academy of Sciences, Beijing, China, in 2020.He is an associate professor at the School of Information Engineering, Ningxia University, and serves as director of the Cyberspace Security Institute at the Ningxia Key Laboratory of Artificial Intelligence and Information Security for Channeling Computing Resources from the East to the West. His research interests include cryptographic application security, system security, blockchain application security, and network public opinion analysis.

E-mail: maziqiang@nxu.edu.cn



Jie Guo is a master's student at the School of Information Engineering, Ningxia University, Yinchuan, China. Her research focuses on malicious encrypted traffic detection and network security, particularly on identifying concealed malicious behavior in encrypted network traffic. She also investigates methods for improving the robustness and reliability of network threat-detection systems in practice.

E-mail: 18795241649@163.com