FROM ABNORMAL VISUAL TOKENS TO FAITHFUL DECODING: REBALANCING VISUAL EVIDENCE FOR HALLUCINATION MITIGATION IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) remain vulnerable to hallucinations despite their strong visual understanding and generation capabilities. Existing hallucination mitigation methods mainly focus on language-side biases, while abnormal visual behaviors and corresponding theoretical implications for faithful decoding remain largely underexplored. Through extensive and statistically significant investigations into the causes of hallucination, we reveal that unusually high-magnitude visual tokens exhibit persistent off-target responses and increasingly interfere with grounded object prediction in deeper layers, accompanied by a progressive weakening of visual evidence during decoding. Based on these observations, we propose a Visual Evidence Rebalancing and Grounding Enhancement method called VERGE, a training-free framework that combines evidence-guided channel correction, uncertainty-triggered visual retracing, and prior-calibrated contrastive decoding. Furthermore, we provide theoretical analyses that characterize the improvement of faithful-token preference and the benefit of recovering target-relevant visual information. Extensive comparative experiments demonstrate that VERGE effectively suppresses hallucinations while maintaining strong generation capabilities.

1 INTRODUCTION

MLLMS have recently demonstrated remarkable capabilities in integrating visual perception with the reasoning abilities of Large Language Models, leading to substantial progress in visual question answering, multimodal reasoning, autonomous systems, and open-ended visual interaction (Huang et al., 2023; Wang et al., 2024a; Sima et al., 2024; Lin et al., 2025). Through large-scale vision-language alignment, modern MLLMs can generate semantically rich responses conditioned on complex visual inputs (Achiam et al., 2023; Li et al., 2025a). Despite these advances, they remain prone to hallucination, where generated content appears plausible yet is inconsistent with the actual image, including nonexistent objects, incorrect attributes, and erroneous relations Ji et al. (2023); Liu et al. (2024a). Such failures substantially undermine the reliability of MLLMs, particularly in applications where predictions are expected to remain faithfully grounded in visual evidence.

To mitigate hallucination, a broad range of approaches has been developed. Knowledge-enhanced methods improve generation reliability through external databases, auxiliary expert models, or additional supervision (Zhu et al., 2023a; Qu et al., 2025). Attention and hidden-state intervention methods directly modify intermediate computation during inference to strengthen visual grounding (Liu et al., 2024c; Zou et al., 2025). Decoding-based approaches constrain the generation process or adjust token probabilities to reduce unsupported predictions (Huang et al., 2024), while contrastive learning methods compare alternative visual or textual conditions and favor candidates with stronger visual support (Leng et al., 2024; Wang et al., 2024c). More recently, training-free inference-time approaches have attracted increasing attention since they can be readily applied to pretrained MLLMs without additional parameter optimization and resource cost.

Despite substantial progress in hallucination mitigation, several important questions remain unresolved. First, the visual-side origin of hallucination remains insufficiently understood (Chen et al., 2024a). Existing studies have extensively examined language priors and text-dominated generation, whereas the internal behavior of visual representations has received considerably less systematic investigation. In particular, it remains unclear whether visual tokens exhibit recurrent abnormal patterns, how such patterns affect visual attention and object-level prediction, and whether these re-

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

lationships persist across large-scale decoding events. Second, existing interventions often emphasize suppressing unreliable responses while paying less attention to the preservation and recovery of useful information (Jiang et al., 2025). Aggressive suppression, such as directly down-weighting attention responses, may reduce hallucination but can simultaneously impair the global task understanding, leading to degraded object recall or descriptive coverage. Effective mitigation should jointly remove hallucination-prone signals and recover reliable image-specific evidence to better balance generation diversity and prediction accuracy, rather than relying on behavior-suppression alone. Third, the architectural design of many training-free methods remains largely empirical (Li et al., 2025b). Intervention directions and degrees are typically selected through heuristic observations and parameter tuning, with limited theoretical understanding of how these interventions alter token selection. In particular, it remains unclear whether a given correction can consistently increase the preference for visually faithful tokens over unsupported alternatives. As a result, such methods often exhibit weak interpretability and practical applicability.

To address these issues, we conduct an extensive investigation into the internal visual dynamics associated with hallucination. We identify a subset of high-magnitude visual tokens that attract attention away from object-discriminative regions and become increasingly associated with hallucinated object prediction in deeper layers. We further observe that visual grounding progressively weakens across decoder layers while predictive uncertainty increases, suggesting that the reliability of visual evidence varies substantially throughout generation. Building on these findings, we propose a training-free inference-time framework VERGE that integrates representation correction, visual evidence recovery, and language-prior calibration. We first suppress unreliable high-magnitude visual responses while enhancing informative evidence through evidence-guided asymmetric channel correction. When visual grounding weakens, uncertainty-triggered visual retracing selectively restores image-specific information. Finally, text-prior contrastive decoding reduces the influence of language-dominated candidates while preserving visually plausible alternatives. Beyond empirical effectiveness, we provide theoretical analyses showing that channel-wise correction can enlarge the faithful-token prediction margin under explicit semantic-alignment and local-smoothness conditions, while visual retracing yields a tighter information-sensitive generalization guarantee when the recovered target-relevant information outweighs the induced representation complexity. The primary contributions of this work are summarized as follows:

- We uncover previously underexplored visual behaviors associated with hallucination in MLLMs. Extensive statistical and intervention analyses reveal that high-magnitude visual tokens exhibit abnormal responses, increasingly interfere with grounded object prediction in deeper layers, and coincide with progressive degradation of visual grounding.
- We establish theoretical guarantees for visual evidence rebalancing and recovery. Our analysis shows that channel-wise correction enlarges the prediction margin of visually faithful tokens, while visual retracing improves generalization by recovering target-relevant information under controlled representation complexity.
- We propose VERGE, a training-free method for visual evidence rebalancing and grounding enhancement. Extensive experiments across multiple MLLMs and benchmarks demonstrate effective hallucination mitigation while preserving strong generation capabilities.

2 Case Study into the Origins of Hallucination

2.1 Preliminaries

Given an image v and a textual prefix x, a MLLM with L transformer layers autoregressively generates an output sequence $y = (y_1, \ldots, y_T)$, where each token y_t is generated conditioned on the image, the textual prefix, and the previously generated tokens $y_{<t}$. At layer l, let $h \in \mathbb{R}^d$ denote the hidden state after the attention module. Then, we denote N_v visual representations by $z_1, \ldots, z_{N_v} \in \mathbb{R}^d$, and collect them into the matrix $Z = [z_1, \ldots, z_{N_v}]^\top \in \mathbb{R}^{N_v \times d}$. For a transformer block equipped with a gated feed-forward network FFN, the intermediate activation and FFN output are given by

$$a = \phi(W_{\text{gate}} h) \odot (W_{\text{up}} h), \qquad f = W_{\text{down}} a, \tag{1}$$

where $W_{\text{gate}}, W_{\text{up}} \in \mathbb{R}^{d' \times d}$, $W_{\text{down}} \in \mathbb{R}^{d \times d'}$, and ϕ is the nonlinearity function. Accordingly, $a \in \mathbb{R}^{d'}$ represents the channel-wise intermediate FFN state before the down-projection. Our fine-grained correction operates directly on a, allowing individual FFN channels to be selectively modulated before their states are projected back to the residual stream.

2.2 Empirical Findings and Observations

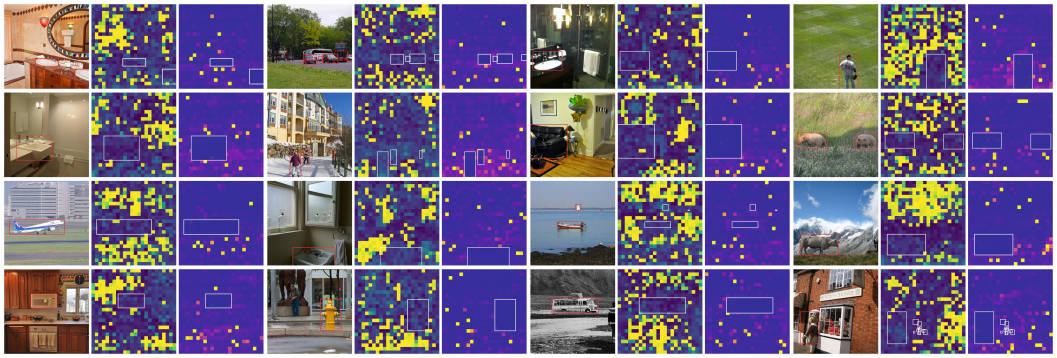


Figure 1: Each triptych shows the COCO image, visual-token norms, and visual attention from the first subword token of a grounded object mention to the visual tokens. Warmer colors indicate larger norms or stronger visual attention.

In this section, we investigate how abnormal visual-token behaviors may eventually contribute to hallucination. Inspired by prior observations that Vision Transformers can spontaneously develop atypical tokens with distinct internal roles (Darcet et al., 2024), we investigate whether atypical visual-token behaviors also arise in MLLMs and disrupt object-level visual grounding.

Irregular Visual Tokens. Using LLaVA-1.5-7B as a representative MLLM, we examine the spatial behavior of projected visual tokens $z_1, \ldots, z_{N_v}$ during object prediction. As shown in Fig. 1, a small subset of visual tokens exhibits unusually large magnitudes, measured by $r_i = \|z_i\|_2$. For each target object, we consider the decoding step $t^\star$ corresponding to its first subword token. Regrading the final decoder layer, $A_{t^\star}$ represents the head-averaged attention matrix, with $at^\star$ indicating the position of the currently analyzed token. The attention over the visual-token positions $\mathcal{V}$ is then given by $A_{t^\star}(a_{t^\star}, \mathcal{V})$. As illustrated in Fig. 1, high-magnitude tokens frequently lie outside object-discriminative regions while still receiving substantial visual attention, resulting in a more diffuse attention pattern. This spatial misalignment suggests that unusually strong visual responses can divert attention from target-relevant evidence, thereby weakening fine-grained visual grounding.

Statistical Characterization of High-Magnitude Visual Tokens. We further identify recurrent high-norm positions by selecting the top 5% spatial positions according to how frequently they exhibit elevated norms across images. As shown in Fig. 2 (left), these positions display a pronounced high-norm tail, revealing a stable model-intrinsic pattern of abnormally strong visual responses. Besides, we stratify the samples by attention dispersion, with the lower and upper 40% forming the low- and high-dispersion groups. The attention dispersion is quantified using the normalized entropy $D_{\text{attn}} = -\frac{1}{\log N_v} \sum_{i \in \mathcal{V}} \widetilde{A}_{t^\star}(a_{t^\star}, \mathcal{V}) \log \widetilde{A}_{t^\star}(a_{t^\star}, \mathcal{V})$, where $\widetilde{A}_{t^\star}$ denotes the normalized attention. High-dispersion samples show a clear shift toward larger visual-token magnitudes compared with low-dispersion samples. Moreover, Fig. 2 (right) reveals a positive association between D_{attn} and the attention mass assigned to the Top 5% high-magnitude candidates. These results demonstrate a consistent coupling between high-magnitude visual responses and abnormal attention allocation.

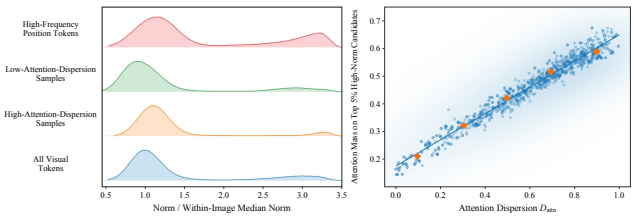


Figure 2: Left: Distributions of image-wise normalized visual-token magnitudes for diverse tokens. Right: Relationship between visual-attention dispersion and the attention mass assigned to the high-magnitude visual tokens.

Functional Characterization of High-Magnitude Visual Tokens. We next examine whether high-magnitude visual tokens play distinct functional roles during decoding. As shown in Fig. 3 (left), the High-Norm Alignment Gap, which measures the relative alignment of textual representations with high-norm versus remaining visual tokens, exhibits clear category-dependent dynamics. Grounded object and function tokens remain negatively aligned, whereas hallucinated object tokens become increasingly aligned with high-norm representations in the middle and later layers. We further con-

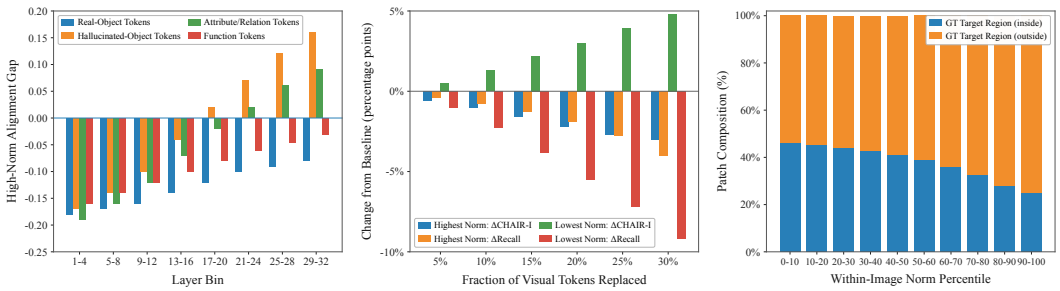


Figure 3: Left: Layer-wise High-Norm Alignment Gap across token categories. Middle: Changes in hallucination rate and object recall under visual-token replacement. Right: Spatial composition of visual tokens across within-image norm percentiles.

duct a visual-token replacement intervention. As shown in Fig. 3 (middle), replacing higher-norm tokens with the global visual-token centroid progressively reduces the instance-level hallucination rate C_I with only a moderate loss in object recall, while replacing lower-norm tokens increases C_I and causes substantially larger recall degradation. Fig. 3 (right) further shows that the fraction of tokens located inside ground-truth object regions decreases toward higher norm percentiles. Together, these results indicate that high-magnitude visual tokens form functionally distinct representations that increasingly bias late-stage decoding away from grounded visual evidence.

Visual Grounding Evolution. We examine the layer-wise evolution of visual grounding from two complementary perspectives: predictive uncertainty and reliance on visual evidence. Predictive uncertainty is quantified by using the entropy of the next-token distribution over the top-K most probable candidates:

$$u = -\frac{1}{\log K} \sum_{k=1}^{K} q_k \log q_k, \qquad (2)$$

where $K = 10$, and q_k denotes the probability of the k-th candidate obtained by projecting the residual hidden state through the language-model head. As shown in Fig. 4 (top-left), uncertainty remains low in early layers but increases markedly in the middle and later layers. In contrast, the visual attention mass in Fig. 4 (top-right) progressively decreases with depth, with stronger attenuation at high-uncertainty positions. This opposing trend suggests that weakened visual reliance in deeper layers is closely associated with less stable predictions, reflecting the progressive emergence of hallucination.

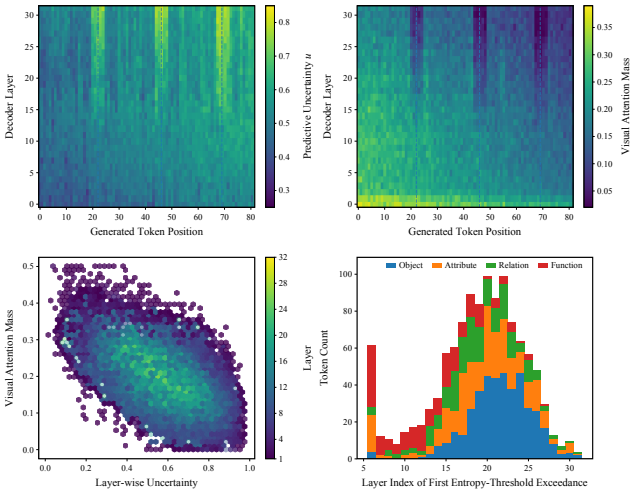


Figure 4: Top: Layer-wise predictive uncertainty and visual-attention mass. Bottom: Inverse uncertainty–attention relationship and adaptive trigger-layer distribution.

3 METHOD

The preceding analysis of hallucination mechanisms suggests that effective mitigation should jointly rectify unreliable visual representations, recover image-specific evidence when grounding weakens, and curb residual language-prior dominance. Accordingly, we develop VERGE, a training-free inference-time framework that unifies asymmetric channel correction, uncertainty-guided visual reinforcement, and language-prior-calibrated contrastive decoding.

3.1 EVIDENCE-GUIDED ASYMMETRIC CHANNEL CORRECTION

Guided by Section 2.2, we can find that visual tokens with unusually large L_2 magnitudes contribute less to fine-grained object-level evidence. Thus, we first partition the visual tokens into a high-magnitude subset and an evidence subset. Given a partition ratio $\rho \in (0, 1)$, we set the number of

high-magnitude tokens as $M = \max(1, \lceil \rho N_v \rceil)$. Let $r_{(1)} \geq r_{(2)} \geq \cdots \geq r_{(N_v)}$ denote the last-layer token magnitudes in descending order. Then, we define the partition threshold as $\theta = r_{(M)}$, i.e., the M-th largest magnitude. Accordingly, the M tokens with the largest magnitudes are assigned to the high-magnitude subset R, while the remaining tokens form the evidence subset E:

$$R = \{i : r_i \geq \theta\}, \qquad E = \{1, \ldots, N_v\} \setminus R. \tag{3}$$

This distinction provides a principled basis for selectively attenuating high-magnitude visual signals while reinforcing effective visual evidence. At each decoding step, we compute the cosine similarity between the current hidden state h and each visual token:

$$s_i = \langle \bar{z}_i, \bar{h} \rangle, \quad \bar{z}_i = \frac{z_i}{\|z_i\|_2}, \ \bar{h} = \frac{h}{\|h\|_2}. \tag{4}$$

Within each subset $S \in \{R, E\}$, we apply a temperature-scaled softmax to s_i, and denote the resulting normalized relevance weight of token i by π_i^S, The corresponding pool-specific visual prototype is then computed as $c^S = \sum_{i \in S} \pi_i^S z_i$. This state-aware formulation preserves task-relevant visual cues more effectively than aggregating all visual tokens into a single static representation. The FFN gate weight W_{gate} consists of d' row vectors $\{w_j\}_{j=1}^{d'}$, each corresponding to an intermediate FFN channel. To characterize the visual preference of each channel, we measure the similarity between its corresponding gate-weight vector and the high-magnitude and evidence prototypes as below:

$$b_j^R = \langle \bar{w}_j, \bar{c}^R \rangle, \qquad b_j^E = \langle \bar{w}_j, \bar{c}^E \rangle, \qquad j = 1, \ldots, d', \tag{5}$$

where $\bar{w}_j = w_j / \|w_j\|_2$ and $\bar{c}^S = c^S / \|c^S\|_2$. A large b_j^E suggests that channel j preferentially captures evidence-bearing visual information and thus merits enhancement. Conversely, a large a_j^R indicates stronger sensitivity to high-magnitude visual features and calls for suppressing the corresponding channel contribution. Building on these channel-wise scores, we introduce an asymmetric correction scheme for the FFN intermediate state a. Specifically, we use multiplicative scaling for suppression and additive adjustment for enhancement, allowing the two correction directions to be controlled separately. We compute a channel-wise suppression mask and enhancement bias as

$$m = \text{clip}\big(1 - \beta\, \sigma(b^R),\ c_{\min},\ 1\big), \qquad g = \alpha\, \text{ReLU}(b^E), \tag{6}$$

where σ denotes the Sigmoid function, $\text{ReLU}(x) = \max(0, x)$, $\alpha, \beta \geq 0$ control the enhancement and suppression strengths, respectively, and $c_{\min} \in [0, 1)$ sets a lower bound on the suppression factor. The sigmoid ensures smooth suppression, while the ReLU retains only positive evidence-related scores for enhancement. The corrected intermediate state is then given by

$$\tilde{a} = a \odot m + g, \qquad \tilde{f} = W_{\text{down}}\, \tilde{a}. \tag{7}$$

The correction is applied to layers $l \in [L_s, L_e]$, while earlier and later layers remain unchanged to retain pure visual representations and preserve linguistic coherence.

To avoid over-correction when the model is already confident, we adapt the correction strength according to its predictive uncertainty. Accordingly, $u \in [0, 1]$ with larger values indicates higher predictive uncertainty. Then, we modulate the fine-grained correction strengths using a mixing coefficient $\eta \in [0, 1]$:

$$\alpha_{\text{eff}} = \alpha \cdot \big[(1 - \eta) + \eta \cdot u\big], \qquad \beta_{\text{eff}} = \beta \cdot \big[(1 - \eta) + \eta \cdot u\big]. \tag{8}$$

As η increases, the correction becomes progressively uncertainty-aware, i.e., confident predictions receive milder intervention, whereas uncertain states retain stronger visual correction.

3.2 Image-Specific Visual Information Reinforcement

The preceding correction mitigates erroneous prior dominance by suppressing unreliable channel responses. However, it only operates on visual information already encoded in the hidden state and cannot directly recover image-specific evidence once visual grounding becomes insufficient. To address this limitation, we introduce an adaptive visual reinforcement mechanism that revisits the original visual tokens when such degradation emerges. To determine when reinforcement should be activated, we examine whether predictive uncertainty can serve as a reliable indicator of weakened visual grounding. As shown in Fig. 4 (bottom-left), token-layer statistics reveal a clear inverse relationship between uncertainty u and visual attention mass, with higher uncertainty

generally accompanied by weaker visual reliance. This observation supports using u to detect insufficient grounding. Fig. 4 (bottom-right) further shows that the first layer exceeding the threshold u_0 varies substantially across generated tokens and semantic categories, with most triggers occurring in the middle and later layers. This variability indicates that a fixed intervention layer is insufficient, motivating a dynamically determined trigger position. Accordingly, the trigger layer is determined as $l^* = \min\{l \mid L_s \le l \le L_e, u > u_0\}$. Once the criterion is first satisfied, the visual adapter is activated at layer $l^* + 1$ to perform a one-shot retrieval through a temporary key-value memory. To match the numerical scale of the visual memory with the corresponding FFN projections, we compute two scaling coefficients as

$$s_1 = \frac{\frac{1}{d'd}\sum_{i=1}^{d'}\sum_{j=1}^{d}|\,(W_{\text{up}})_{ij}\,|}{\frac{1}{N_v d}\sum_{i=1}^{N_v}\sum_{j=1}^{d}|Z_{ij}| + \varepsilon}, \quad s_2 = \frac{\frac{1}{dd'}\sum_{i=1}^{d}\sum_{j=1}^{d'}|\,(W_{\text{down}})_{ij}\,|}{\frac{1}{N_v d}\sum_{i=1}^{N_v}\sum_{j=1}^{d}|Z_{ij}| + \varepsilon}, \tag{9}$$

where ε ensures numerical stability. The visual key and value matrices are then constructed by rescaling the original visual tokens, i.e., $K = s_1 Z \in \mathbb{R}^{N_v \times d}$ and $V = s_2 Z \in \mathbb{R}^{N_v \times d}$. Given the hidden state h at layer $l^* + 1$, the retrieved visual representation is computed as $\Delta = \left(hK^\top\right) V \in \mathbb{R}^d$. The relevance scores $hK^\top \in \mathbb{R}^{N_v}$ measure the correspondence between the current hidden state and individual visual tokens, which are used to aggregate the associated visual values. In this way, relevant visual evidence can be selectively emphasized. Before incorporating the retrieved visual evidence into the FFN pathway, we further align its magnitude with the corrected FFN output:

$$\overline{\Delta} = \frac{\frac{1}{d}\sum_{j=1}^{d}\left|\tilde{f}_j\right|}{\frac{1}{d}\sum_{j=1}^{d}|\Delta_j| + \varepsilon}\Delta, \quad \hat{f} = (1-\lambda)\tilde{f} + \lambda\overline{\Delta}, \tag{10}$$

where $\lambda \in [0, 1]$ controls the contribution of the added visual evidence.

3.3 LANGUAGE-PRIOR-CALIBRATED CONTRASTIVE DECODING

Hidden-space correction strengthens the visual grounding of the generated representation, yet the final prediction distribution can still be dominated by statistical language priors. In particular, tokens strongly favored by the textual context may receive excessive prediction probability despite limited visual support. Therefore, we introduce a vocabulary-level contrastive decoding scheme that explicitly separates visual evidence from language prior preference. At each decoding step, the image-conditioned branch produces the log-probability distribution $\ell^+ = \log p_\theta(y_t \mid v, x, y_{<t})$, while a parallel text-only branch without the image yields $\ell^- = \log p_\theta(y_t \mid \varnothing, x, y_{<t})$. The text-only distribution provides a direct estimate of the language prior without introducing perturbations to the visual input. Then, we construct the contrastive score as $\tilde{\ell} = (1+\delta)\,\ell^+ - \delta\,\ell^-$, where $\delta \ge 0$ controls the contrastive strength. This formulation favors tokens whose likelihood increases under visual conditioning while suppressing tokens already strongly supported by the text-only branch. However, contrastive reweighting may also overemphasize tokens with a large relative visual gain despite their relatively low probability under the image-conditioned model. To prevent such implausible candidates from being promoted, we impose an adaptive plausibility constraint:

$$\kappa(\delta) = \kappa_0 \cdot \min\Big(1,\ \big(\delta/\delta_{\text{ref}}\big)^q\Big), \tag{11}$$

$$\mathcal{V} = \Big\{\, w \;\Big|\; p_\theta(w \mid v, x, y_{<t}) \ge \kappa(\delta) \cdot \max_{w'} p_\theta(w' \mid v, x, y_{<t}) \,\Big\}. \tag{12}$$

Here, $\kappa_0 \in (0, 1]$ denotes the base plausibility ratio, δ_{ref} is the reference contrastive strength, and $q > 0$ controls the tightening rate of the constraint. The final token distribution is given by

$$p(y_t = w) \;\propto\; 1[w \in \mathcal{V}] \cdot \exp\big(\tilde{\ell}(w)\big). \tag{13}$$

This constraint preserves visually supported candidates while excluding tokens that remain insufficiently plausible under the image-guided branch.

4 THEORETICAL ANALYSIS

In this section, we provide theoretical guarantees for the two visual-side interventions in VERGE, with complete derivations and further explanations provided in Appendices B, C and D.

Theorem 1 (Faithful-margin amplification by channel-wise correction) *Consider the channel-wise correction $\widetilde{a} = a + r(u)d_0$, where a is the FFN intermediate activation, d_0 is the correction direction, and $r(u) > 0$ is the uncertainty-dependent correction strength. For the visually faithful token $y^\star$ and any competitor $y \neq y^\star$, let $M_y(a)$ denote their pairwise prediction margin, and define the faithful margin as $\Delta(a) = \min_{y \neq y^\star} M_y(a)$. Assume that the correction direction is locally aligned with faithful prediction, with a uniform first-order margin gain lower bounded by $A_\star > 0$, and that the pairwise margins are locally smooth along the correction path, with curvature uniformly bounded by $L_\star > 0$. Defining $D = \|d_0\|_2$, the corrected faithful margin then satisfies*

$$\Delta(\widetilde{a}) \geq \Delta(a) + \xi(u), \qquad \xi(u) = r(u)A_\star - \frac{L_\star}{2} r(u)^2 D^2. \tag{14}$$

Hence, if $0 < r(u) < \frac{2A_\star}{L_\star D^2}$, then $\xi(u) > 0$. Let $\epsilon_{\mathrm{cal}}(\xi_0)$ denote the probability that the correction fails to guarantee a margin increase of at least ξ_0, and let s_{base} and s_{corr} denote the predictors before and after channel-wise correction, respectively. For $s \in \{s_{\mathrm{base}}, s_{\mathrm{corr}}\}$, let $\Delta_s(x, y^\star)$ denote the faithful prediction margin for target token $y^\star$. Then, we define the population token-selection risk as $R(s) = \mathbb{P}[\Delta_s(x, y^\star) \leq 0]$ and the empirical low-margin frequency at level γ as $\widehat{F}_s(\gamma) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}[\Delta_s(x_i, y_i^\star) \leq \gamma]$. Suppose that the inference-time configurations form a finite family $\mathcal{H}$ whose induced margins are uniformly bounded by B_0. With probability at least $1-\delta$ over n decoding contexts, we have

$$R(s_{\mathrm{corr}}) \leq \widehat{F}_{s_{\mathrm{base}}}(\gamma - \xi_0) + \widehat{\epsilon}_{\mathrm{cal}}(\xi_0) + \Omega_{s_{\mathrm{corr}}}(n, \gamma, \delta), \tag{15}$$

where

$$\Omega_{s_{\mathrm{corr}}} = \frac{2B_0}{\gamma}\sqrt{\frac{2\log|\mathcal{H}|}{n}} + 3\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{16}$$

Accordingly, the corrected predictor admits a tighter margin-based population guarantee whenever

$$\widehat{F}_{s_{\mathrm{base}}}(\gamma) - \widehat{F}_{s_{\mathrm{base}}}(\gamma - \xi_0) > \widehat{\epsilon}_{\mathrm{cal}}(\xi_0) + \Omega_{s_{\mathrm{corr}}} - \Omega_{s_{\mathrm{base}}}. \tag{17}$$

Theorem 2 (Generalization Guarantee for Visual Retracing) *Let the uncertainty-trigger indicator denoted by $T \in \{0, 1\}$, and consider the triggered distribution conditioned on $T = 1$. Besides, the visually faithful next-token label is defined as $Y \in \{1, \ldots, K\}$, where K denotes the vocabulary size. Let S_0 and S_1 denote the representations before and after visual retracing, respectively, and define the recovered target-information gain as $G_{\mathrm{VR}} = I(S_1; Y \mid T = 1) - I(S_0; Y \mid T = 1)$. Assume that the loss ℓ is bounded and Lipschitz, the frozen suffix mapping ψ is Lipschitz on a common activation envelope containing S_0 and S_1, and the class-conditional self-information has bounded coordinate-wise sensitivity c_S over m nuisance factors. Then, we define*

$$\widehat{\ell}_S(s, y) = \ell(\psi(0), y) + L_\ell L_\psi \|s\|_2, \qquad \rho_S = \frac{1}{n}\sum_{i=1}^{n}\Big[\widehat{\ell}_S(S_i, Y_i) - \ell(\psi(S_i), Y_i)\Big], \tag{18}$$

and $M_S = \sup_{s,y} \widehat{\ell}_S(s, y)$ on the same envelope. For $\gamma, \delta \in (0, 1)$, we define $\epsilon_S = c_S\sqrt{\frac{m}{2}\log\frac{\sqrt{n}}{\gamma}}$, and $\Gamma_S = \epsilon_S + \log\frac{2K}{\delta}$. With $\delta_1 = \delta_2 = \delta/2$, let $A_1 = M_S$, $A_2 = M_S\sqrt{2K}$, $A_3 = 2\sqrt{2K}\, M_S$, and $A_4 = \frac{2KM_S}{3}$. With probability at least $1 - \delta$ over n triggered contexts, every admissible representation S satisfies

$$\mathrm{Gen}_n(S) \leq \rho_S + A_1\frac{\gamma}{\sqrt{n}} + A_2\frac{\sqrt{\gamma\log(4K/\delta)}}{n^{3/4}} + A_4\frac{\log(4K/\delta)}{n} + A_3\sqrt{\frac{H(S \mid Y, T = 1) + \Gamma_S}{n}}. \tag{19}$$

For $j \in \{0, 1\}$, we define $V_j = H(S_j \mid T = 1) - I(S_j; Y \mid T = 1) + \Gamma_{S_j}$, $\Delta H = H(S_1 \mid T = 1) - H(S_0 \mid T = 1)$, and $\Delta\Gamma = \Gamma_{S_1} - \Gamma_{S_0}$. The representation-dependent complexity term is strictly reduced after visual retracing whenever $G_{\mathrm{VR}} > \Delta H + \Delta\Gamma$. Moreover, the complete generalization upper bound is strictly tighter for S_1 than for S_0 whenever

$$\rho_{S_1} - \rho_{S_0} < \frac{A_3}{\sqrt{n}}\frac{G_{\mathrm{VR}} - \Delta H - \Delta\Gamma}{\sqrt{V_0} + \sqrt{V_1}}. \tag{20}$$

5 EXPERIMENT

In this section, we present the basic experimental settings and analyze the main comparative results, while additional experimental results, ablation studies, and parameter analyses are provided in Appendices H.1, H.2, and H.3, respectively.

Table 1: Performance evaluation on POPE benchmark. The best results are highlighted in bold.

Methods	LLaVA-1.5-7B		LLaVA-1.5-13B		Qwen2.5-VL-7B		InternVL3-8B		GLM-4V-9B	
	Accuracy ↑	F1 ↑	Accuracy ↑	F1 ↑	Accuracy ↑	F1 ↑	Accuracy ↑	F1 ↑	Accuracy ↑	F1 ↑
Regular	84.1 ↑0.0	82.8 ↑0.0	87.0 ↑0.0	86.3 ↑0.0	87.7 ↑0.0	86.3 ↑0.0	90.8 ↑0.0	90.6 ↑0.0	89.3 ↑0.0	89.0 ↑0.0
ICD	85.5 ↑1.4	84.7 ↑1.9	87.2 ↑0.2	86.5 ↑0.2	88.0 ↑0.3	87.0 ↑0.7	90.8 ↑0.0	90.6 ↑0.0	89.8 ↑0.5	89.5 ↑0.5
VCD	86.0 ↑1.9	85.2 ↑2.4	87.1 ↑0.1	86.4 ↑0.1	88.4 ↑0.7	87.4 ↑1.1	90.6 ↓0.2	90.6 ↑0.0	89.9 ↑0.6	89.7 ↑0.7
PAI	86.2 ↑2.1	85.4 ↑2.6	86.2 ↓0.8	84.8 ↓1.5	88.3 ↑0.6	87.2 ↑0.9	90.7 ↓0.1	90.5 ↓0.1	89.6 ↑0.3	89.3 ↑0.3
Devils	78.1 ↓6.0	80.3 ↓2.5	85.3 ↓1.7	85.1 ↓1.2	85.5 ↓2.2	83.4 ↓2.9	90.5 ↓0.3	90.2 ↓0.4	89.5 ↑0.2	89.2 ↑0.2
MemVR	86.8 ↑2.7	85.9 ↑3.1	88.4 ↑1.4	88.0 ↑1.7	88.6 ↑0.9	87.6 ↑1.3	90.9 ↑0.1	**90.7** ↑0.1	89.9 ↑0.6	89.7 ↑0.7
VISTA	87.2 ↑3.1	86.5 ↑3.7	88.6 ↑1.6	88.3 ↑2.0	87.9 ↑0.2	86.7 ↑0.4	90.4 ↓0.4	90.1 ↓0.5	90.0 ↑0.7	89.8 ↑0.8
Ours	**87.4** ↑3.3	**86.7** ↑3.9	**88.7** ↑1.7	**88.5** ↑2.2	**89.1** ↑1.4	**88.3** ↑2.0	**91.0** ↑0.2	**90.7** ↑0.1	**90.1** ↑0.8	**89.9** ↑0.9

5.1 Experimental Configurations

Model Backbones. We evaluate VERGE on five representative MLLM backbones: LLaVA-1.5-7B, LLaVA-1.5-13B, Qwen2.5-VL-7B, InternVL3-8B, and GLM-4V-9B. The two LLaVA variants enable controlled evaluation across model scales, while Qwen2.5-VL, InternVL3, and GLM-4V diversify the visual encoders, multimodal alignment schemes, and language backbones. This heterogeneous suite provides a rigorous test of VERGE across model scales and architectural paradigms.

Comparison Baselines. We compare VERGE with representative training-free hallucination mitigation methods spanning complementary inference-time paradigms. OPERA regulates decoding via over-trust penalties and retrospective reallocation; VCD and ICD perform contrastive decoding with distorted visual inputs and disturbance instructions, respectively. PAI and DeVILs recalibrate visual attention at the token and head levels, while MemVR and VISTA recover latent visual evidence through memory-space retracing and activation-space steering, respectively.

Benchmarks and Implementation Details. We evaluate both hallucination mitigation and general multimodal capability. For hallucination evaluation, we adopt CHAIR, POPE, and HallusionBench. CHAIR measures object hallucination in free-form descriptions, POPE evaluates object-existence hallucination through balanced polling, and HallusionBench assess broader hallucination behaviors across challenging visual scenarios. To examine capability preservation, we further evaluate on LLaVA-Bench, MME, MM-Vet, VizWiz, and MMBench, covering open-ended generation, visual perception, multimodal reasoning, real-world visual question answering, and general multimodal understanding. We adopt the original model with standard decoding as the Greedy baseline and follow the official implementations and recommended inference configurations of all competing methods. The parameter configurations and tuning procedures of our method are detailed in Appendix H.3. All experiments were conducted using NVIDIA A100 PCIe GPUs with 80 GB of memory.

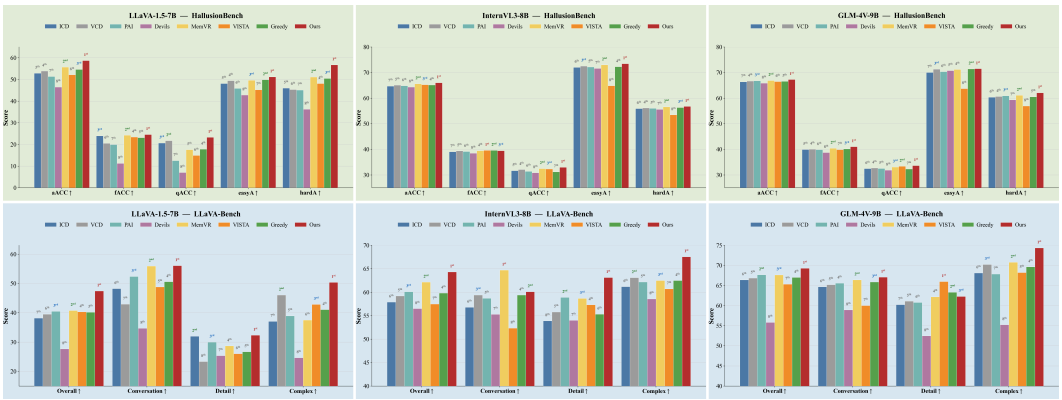


Figure 5: HallusionBench and LLaVA-Bench evaluation results of different methods.

5.2 Results of Comparison Experiments

Results on Hallucination-oriented Benchmarks. As shown in Tables 1, On POPE, VERGE attains the highest accuracy across all backbones and the best or tied-best F1 scores, yielding consistent gains on both relatively hallucination-prone models and already strong backbones. In contrast, comparison methods exhibit varying degrees of model-dependent fluctuation and even degradation on certain backbones. The consistent improvements of VERGE across heterogeneous architectures and complementary hallucination protocols demonstrate its robust generalizability, suggesting that coordinated visual-evidence rebalancing and visual-information retracing provide more reliable ground-

Table 2: Performance test on MME Hallucination subset. The best results are highlighted in bold.

Model	Methods	MME-Hall	Object-Level		Attribute-Level		Model	Methods	MME-Hall	Object-Level		Attribute-Level	
		Total ↑	Existence ↑	Count ↑	Position ↑	Color ↑			Total ↑	Existence ↑	Count ↑	Position ↑	Color ↑
LLaVA-1.5-7B	Regular	631.7 ↑0.0	**195.0** ↑0.0	158.3 ↑0.0	123.3 ↑0.0	155.0 ↑0.0	InternVL3-8B	Regular	698.3 ↑0.0	195.0 ↑0.0	150.0 ↑0.0	163.3 ↑0.0	190.0 ↑0.0
	ICD	630.0 ↓1.7	190.0 ↓5.0	143.3 ↓15.0	**133.3** ↑10.0	163.3 ↑8.3		ICD	693.3 ↓5.0	195.0 ↑0.0	150.0 ↑0.0	163.3 ↑0.0	185.0 ↓5.0
	VCD	630.0 ↓1.7	190.0 ↓5.0	148.3 ↓10.0	**133.3** ↑10.0	158.3 ↑3.3		VCD	668.3 ↓30.0	195.0 ↑0.0	145.0 ↓5.0	153.3 ↓10.0	175.0 ↓15.0
	PAI	621.7 ↓10.0	**195.0** ↑0.0	148.3 ↓10.0	123.3 ↑0.0	155.0 ↑0.0		PAI	700.0 ↑1.7	**200.0** ↑5.0	**155.0** ↑5.0	**170.0** ↑6.7	175.0 ↓15.0
	Devils	605.0 ↓26.7	**195.0** ↑0.0	148.3 ↓10.0	106.7 ↓16.6	155.0 ↑0.0		Devils	703.3 ↑5.0	195.0 ↑0.0	**155.0** ↑5.0	163.3 ↑0.0	190.0 ↑0.0
	MemVR	648.3 ↑16.6	190.0 ↓5.0	155.0 ↓3.3	**133.3** ↑10.0	**170.0** ↑15.0		MemVR	703.3 ↑5.0	195.0 ↑0.0	150.0 ↑0.0	168.3 ↑5.0	190.0 ↑0.0
	VISTA	638.3 ↑6.6	190.0 ↓5.0	150.0 ↓8.3	**133.3** ↑10.0	165.0 ↑10.0		VISTA	686.7 ↓11.6	190.0 ↓5.0	150.0 ↑0.0	166.7 ↑3.4	180.0 ↓10.0
	Ours	**653.3** ↑21.6	190.0 ↓5.0	**160.0** ↑1.7	**133.3** ↑10.0	**170.0** ↑15.0		Ours	**708.3** ↑10.0	195.0 ↑0.0	150.0 ↑0.0	168.3 ↑5.0	**195.0** ↑5.0
Qwen2.5-VL-7B	Regular	705.0 ↑0.0	195.0 ↑0.0	**160.0** ↑0.0	**160.0** ↑0.0	190.0 ↑0.0	GLM-4V-9B	Regular	**708.3** ↑0.0	**200.0** ↑0.0	**173.3** ↑0.0	156.7 ↑0.0	178.3 ↑0.0
	ICD	705.0 ↑0.0	195.0 ↑0.0	**160.0** ↑0.0	**160.0** ↑0.0	190.0 ↑0.0		ICD	**708.3** ↑0.0	**200.0** ↑0.0	**173.3** ↑0.0	156.7 ↑0.0	178.3 ↑0.0
	VCD	700.0 ↓5.0	195.0 ↑0.0	**160.0** ↑0.0	**160.0** ↑0.0	185.0 ↓5.0		VCD	701.7 ↓6.6	**200.0** ↑0.0	165.0 ↓8.3	**161.7** ↑5.0	175.0 ↓3.3
	PAI	690.0 ↓15.0	190.0 ↓5.0	150.0 ↓10.0	150.0 ↓10.0	**200.0** ↑10.0		PAI	701.7 ↓6.6	**200.0** ↑0.0	**173.3** ↑0.0	143.3 ↓13.4	**185.0** ↑6.7
	Devils	683.3 ↓21.7	190.0 ↓5.0	143.3 ↓16.7	155.0 ↓5.0	195.0 ↑5.0		Devils	695.0 ↓13.3	**200.0** ↑0.0	163.3 ↓10.0	153.3 ↓3.4	178.3 ↑0.0
	MemVR	700.0 ↓5.0	**200.0** ↑5.0	155.0 ↓5.0	155.0 ↓5.0	190.0 ↑0.0		MemVR	693.3 ↓15.0	**200.0** ↑0.0	163.3 ↓10.0	156.7 ↑0.0	173.3 ↓5.0
	VISTA	700.0 ↓5.0	195.0 ↑0.0	155.0 ↓5.0	**160.0** ↑0.0	190.0 ↑0.0		VISTA	695.0 ↓13.3	**200.0** ↑0.0	165.0 ↓8.3	151.7 ↓5.0	178.3 ↑0.0
	Ours	**715.0** ↑10.0	**200.0** ↑5.0	**160.0** ↑0.0	**160.0** ↑0.0	195.0 ↑5.0		Ours	**708.3** ↑0.0	**200.0** ↑0.0	**173.3** ↑0.0	156.7 ↑0.0	178.3 ↑0.0

ing than isolated inference-time interventions. As shown in Fig. 5, on HallusionBench, VERGE achieves consistently strong performance across LLaVA-1.5-7B, InternVL3-8B, and GLM-4V-9B, attaining the best results on most evaluation metrics. The gains are particularly evident on the more challenging qACC and hardA metrics, while maintaining competitive performance on the remaining criteria. In contrast to the metric- and backbone-dependent fluctuations of competing methods, VERGE exhibits a more balanced and consistent improvement, further demonstrating the effectiveness of visual-side correction and restoration in reinforcing faithful visual grounding.

Results on General-purpose Benchmarks. As shown in Fig. 5, VERGE achieves consistently competitive overall performance on LLaVA-Bench, with particularly strong results on the Conversation and Complex dimensions. Compared with VERGE, ICD and VCD generally yield limited improvements over Greedy, while PAI and MemVR exhibit noticeable fluctuations across capability dimensions. Therefore, VERGE maintains a more balanced performance profile, indicating that its hallucination mitigation largely preserves general multimodal reasoning capabilities. Table 2 further evaluates hallucination mitigation on the MME-Hall. VERGE achieves the highest total scores on LLaVA-1.5-7B and Qwen2.5-VL-7B, improving the regular baseline by 21.6 and 10.0 points, respectively, while also reaching the best score of 708.3 on InternVL3-8B and matching the baseline optimum on GLM-4V-9B. At the category level, the gains are distributed across both object- and attribute-level evaluation, with notable improvements in Count, Position and Color. Besides. according to Fig. 6, VERGE maintains consistently strong performance on VizWiz, MM-Vet, and MMBench, ranking first in nearly all evaluated settings. In comparison, methods such as PAI and ICD exhibit pronounced performance degradation on certain backbones, revealing a less stable trade-off between hallucination mitigation and general capability preservation. Taken together, these results demonstrate that VERGE effectively suppresses hallucinations while preserving strong generation and visual understanding capabilities.

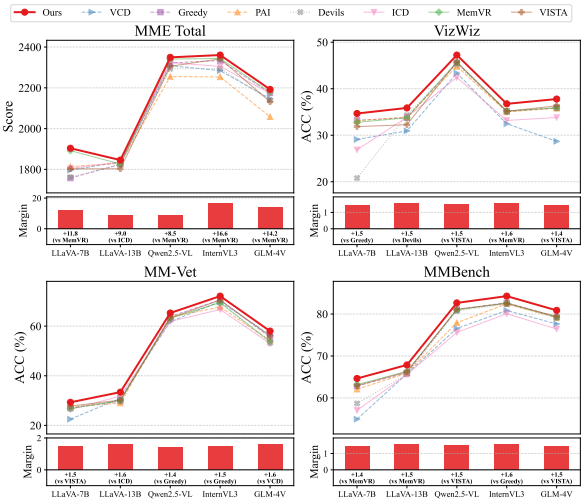


Figure 6: Performance evaluation on four benchmarks. The lower margin panels report the absolute performance gain over the second-best method.

6 Conclusion

We investigate visual-side hallucination mechanisms in MLLMs and reveal that high-magnitude visual tokens increasingly interfere with grounded prediction alongside progressively weakened visual reliance. Building on these findings, we propose VERGE, a training-free framework that combines channel correction, visual retracing, and prior-calibrated decoding. Our theoretical analysis characterizes the benefits of visual evidence rebalancing and recovery. Extensive experiments show that VERGE consistently mitigates hallucination while preserving strong reasoning capabilities.

AI USE STATEMENT

In this work, generative AI tools were used solely to assist with manuscript writing and presentation, including language editing, improving clarity and readability, and organizing experimental results into tables and figures. Generative AI was not used for developing research ideas or methods, formulating scientific claims, designing experiments, analyzing data, interpreting results, conducting literature searches, or deriving theoretical results. All AI-assisted edits and visualizations were carefully reviewed by the authors, who take full responsibility for the final content of the manuscript.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

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

Ali Furkan Biten, Lluis Gomez, and Dimosthenis Karatzas. Let there be a clock on the beach: Reducing object hallucination in image captioning. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 1381–1390, 2022.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. InternVL: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 24185–24198, 2024b.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. A survey on multimodal large language models for autonomous driving. In *IEEE/CVF Winter Conference on Applications of Computer Vision Workshops*, 2024.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *International Conference on Learning Representations*, volume 2024, pp. 2632–2652, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2019.

Danny Driess, Fei Xia, Mehdi S. M. Sajjadi, Corey Lynch, Aakanksha Chowdhery, Brian Ichter, Ayzaan Wahid, Jonathan Tompson, Quan Vuong, Tianhe Yu, Wenlong Huang, Yevgen Chebotar, Pierre Sermanet, Daniel Duckworth, Sergey Levine, Vincent Vanhoucke, Karol Hausman, Marc Toussaint, Klaus Greff, Andy Zeng, Igor Mordatch, and Pete Florence. PaLM-E: An embodied multimodal language model. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 8469–8488. PMLR, 2023.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in neural information processing systems*, 38, 2026.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14375–14385. IEEE, 2024.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. Vizwiz grand challenge: Answering visual questions from blind people. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 3608–3617. IEEE, 2018.

Wenyi Hong, Weihan Wang, Ming Ding, Wenmeng Yu, Qingsong Lv, Yan Wang, Yean Cheng, Shiyu Huang, Junhui Ji, Zhao Xue, et al. CogVLM2: Visual language models for image and video understanding. *arXiv preprint arXiv:2408.16500*, 2024.

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

Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liu, Kriti Aggarwal, Zewen Chi, Nils Johan Bertil Bjorck, Vishrav Chaudhary, Subhojit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. In *Advances in Neural Information Processing Systems*, 2023.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 2023.

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

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Jae Myung Kim, A. Sophia Koepke, Cordelia Schmid, and Zeynep Akata. Exposing and mitigating spurious correlations for cross-modal retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pp. 2585–2595, 2023.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024.

Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. LLaVA-NeXT-Interleave: Tackling multi-image, video, and 3d in large multimodal models. *arXiv preprint arXiv:2407.07895*, 2024.

Liunian Harold Li, Mark Yatskar, Da Yin, Cho-Jui Hsieh, and Kai-Wei Chang. Visualbert: A simple and performant baseline for vision and language. In *Annual Meeting of the Association for Computational Linguistics*, 2019. doi: 10.18653/v1/P19-1431.

Quanjiang Li, Tianxiang Xu, Tingjin Luo, Yan Zhong, Yang Li, Yiyun Zhou, and Chenping Hou. Theory-driven label-specific representation for incomplete multi-view multi-label learning. In *Advances in Neural Information Processing Systems 38*, volume 38, pp. 135155–135196, 2025a. doi: 10.52202/085713-4066.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 292–305. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.emnlp-main.20.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. 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*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35799–35819. PMLR, 2025b.

Tianwei Lin, Wenqiao Zhang, Sijing Li, Yuqian Yuan, Binhe Yu, Haoyuan Li, Wanggui He, Hao Jiang, Mengze Li, Xiaohui Song, Siliang Tang, Jun Xiao, Hui Lin, Yueting Zhuang, and Beng Chin Ooi. Healthgpt: A medical large vision-language model for unifying comprehension and generation via heterogeneous knowledge adaptation. *arXiv preprint arXiv:2502.09838*, 2025.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024b.

Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in LVLMs. In *European Conference on Computer Vision (ECCV) 2024*, pp. 125–140, 2024c.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024d.

Xiaoye Qu, Qiyuan Chen, Wei Wei, Jiashuo Sun, Daizong Liu, and Jianfeng Dong. Alleviating hallucination in large vision-language models with active retrieval augmentation. *ACM Transactions on Multimedia Computing, Communications, and Applications*, 21(9):1–22, 2025. doi: 10.1145/3742434.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045. Association for Computational Linguistics, 2018. doi: 10.18653/v1/D18-1437.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *European Conference on Computer Vision*, 2024.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. VideoBERT: A joint model for video and language representation learning. In *IEEE/CVF International Conference on Computer Vision*, 2019.

Chongjun Tu, Peng Ye, Dongzhan Zhou, Lei Bai, Gang Yu, Tao Chen, and Wanli Ouyang. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of MLLMs. *International Journal of Computer Vision*, 134, 2026. doi: 10.1007/s11263-025-02607-z. Article 22.

Dongsheng Wang, Natraj Raman, Mathieu Sibue, Zhiqiang Ma, Petr Babkin, Simerjot Kaur, Yulong Pei, Armineh Nourbakhsh, and Xiaomo Liu. Docllm: A layout-aware generative language model for multimodal document understanding. In *Annual Meeting of the Association for Computational Linguistics*, 2024a. doi: 10.18653/v1/2024.acl-long.463.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-VL: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024b.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 15840–15853. Association for Computational Linguistics, 2024c. doi: 10.18653/v1/2024.findings-acl.937.

Junfei Wu, Qiang Liu, Ding Wang, Jinghao Zhang, Shu Wu, Liang Wang, and Tieniu Tan. Logical closed loop: Uncovering object hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 6944–6962. Association for Computational Linguistics, 2024. doi: 10.18653/v1/2024.findings-acl.414.

Diji Yang, Jinmeng Rao, Kezhen Chen, Xiaoyuan Guo, Yawen Zhang, Jie Yang, and Yi Zhang. IM-RAG: Multi-round retrieval-augmented generation through learning inner monologues. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pp. 730–740. Association for Computing Machinery, 2024. doi: 10.1145/3626772.3657760.

Qifan Yu, Juncheng Li, Longhui Wei, Liang Pang, Wentao Ye, Bosheng Qin, Siliang Tang, Qi Tian, and Yueting Zhuang. HalluciDoctor: Mitigating hallucinatory toxicity in visual instruction data. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 12944–12953, 2024.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. Mitigating object hallucination in large vision-language models via image-grounded guidance. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 77461–77486. PMLR, 2025.

Beier Zhu, Yulei Niu, Saeil Lee, Minhoe Hur, and Hanwang Zhang. Debiased fine-tuning for vision-language models by prompt regularization. *Proceedings of the AAAI Conference on Artificial Intelligence*, 37(3):3834–3842, 2023a. doi: 10.1609/aaai.v37i3.25496.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023b.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Yuchen Duan, Hao Tian, Weijie Su, Jie Shao, et al. InternVL3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025.

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

A Related Work

A.1 Multimodal Large Language Models

Vision-language modeling has undergone a substantial paradigm shift in recent years. Early vision-language models commonly employed BERT-style encoders (Devlin et al., 2019) to learn joint representations from paired images and text (Li et al., 2019; Sun et al., 2019). These encoder-based models established effective cross-modal alignment, but their task-specific heads and predominantly discriminative objectives limited open-ended generation. The rise of autoregressive large language models (LLMs) (Brown et al., 2020; Achiam et al., 2023) subsequently motivated multimodal large language models (MLLMs) that connect visual encoders to generative language backbones (Alayrac et al., 2022; Driess et al., 2023). Most contemporary MLLMs encode an image into visual features, map or resample those features into the language embedding space, and use the language model to autoregressively predict textual response tokens. Representative systems instantiate this cross-modal interface differently. LLaVA-1.5 (Liu et al., 2024b) uses an MLP projector and visual instruction tuning to align CLIP features (Radford et al., 2021) with its language model. MiniGPT-4 (Zhu et al., 2023b) instead uses a pretrained vision encoder and Q-Former followed by a projection layer to connect visual features to the language model. These architectures provide standard foundations for studying how MLLMs encode and propagate visual evidence during autoregressive generation.

Recent MLLMs further improve resolution handling, spatial grounding, and multimodal reasoning. Qwen-VL (Bai et al., 2023) introduced a generalist model for visual question answering, localization, and text reading, while Qwen2-VL (Wang et al., 2024b) extended the Qwen-VL family to arbitrary image resolutions and long videos. Qwen2.5-VL (Bai et al., 2025) further equips its native dynamic-resolution vision transformer with window attention and absolute time encoding for videos, thereby strengthening localization, document parsing, and long-video understanding. In parallel, InternVL (Chen et al., 2024b) scales both its vision foundation model and its alignment with an LLM. InternVL3 (Zhu et al., 2025) adopts native multimodal pretraining to develop linguistic and multimodal capabilities jointly, and its variable visual position encoding supports extended multimodal contexts. GLM-4V, introduced alongside CogVLM2 (Hong et al., 2024), uses the GLM-4-9B language backbone and the CogVLM2 data and training recipes, but removes the visual-expert modules to reduce model size. LLaVA-NeXT-Interleave (Li et al., 2024) extends visual instruction tuning to interleaved multi-image, video, and three-dimensional inputs. Despite these architectural and data advances, MLLMs still generate fluent responses that lack support from their visual inputs, which indicates that larger perception and language components alone do not resolve hallucination.

A.2 Object Hallucination in MLLMs

Hallucination remains a persistent and consequential challenge in modern MLLMs (Ji et al., 2023; Liu et al., 2024a). Hallucinated vision-language outputs may contain nonexistent objects, incorrect attributes, erroneous counts, or unsupported relations. Object hallucination, the focus of this work, occurs when a model mentions objects that are absent from the input image (Li et al., 2023; Rohrbach et al., 2018). Object hallucinations are especially concerning in safety-critical multimodal systems such as autonomous driving, where a plausible but visually unsupported statement can lead to unsafe downstream decisions (Cui et al., 2024). Prior analyses associate hallucination with several interacting factors, including language priors induced by frequent object co-occurrences, insufficient cross-modal alignment, weakened visual evidence in intermediate representations, and increasing reliance on previously generated text (Biten et al., 2022; Liu et al., 2024c; Jiang et al.,

2025; Zou et al., 2025; Li et al., 2025b). These diverse causes have motivated evaluation protocols that examine complementary aspects of visual faithfulness. CHAIR (Rohrbach et al., 2018) measures hallucinated object mentions in generated captions, whereas POPE (Li et al., 2023) evaluates object presence through balanced polling-based questions. HaELM (Wang et al., 2023) instead trains an LLM-based evaluator to assess hallucination in free-form MLLM responses.

Existing mitigation methods intervene at different stages of model construction and generation. Biten et al. (Biten et al., 2022) reduce object bias in image captioning through sentence-level augmentation that flattens object co-occurrence statistics. Related training-time studies expose spurious correlations in cross-modal retrieval (Kim et al., 2023), apply hallucination-augmented contrastive learning (Jiang et al., 2024), clean hallucinatory visual instruction data (Yu et al., 2024), or regularize prompts during debiased fine-tuning (Zhu et al., 2023a). Retrieval-oriented methods add external evidence through active retrieval for LVLMs (Qu et al., 2025); broader LLM research also supports multi-round retrieval through learned inner monologues (Yang et al., 2024). MARINE (Zhao et al., 2025) obtains object-level guidance from open-source vision models, while LogicCheckGPT (Wu et al., 2024) uses a logical closed loop of follow-up questions to expose inconsistent object claims. Depending on their design, these methods introduce training data, parameter updates, retrieval infrastructure, auxiliary models, or additional verification calls. These dependencies can reduce portability across model families and complicate deployment. Consequently, a growing body of work studies training-free mitigation that modifies decoding distributions, attention weights, or hidden representations during inference. We organize these training-free approaches below according to the stage at which each method restores visual grounding.

A.3 Decoding-Time Hallucination Mitigation

One stream of training-free research modifies candidate selection during decoding so that visually grounded tokens receive greater probability without parameter updates. OPERA (Huang et al., 2024) associates object hallucination with an over-trust pattern in which the model concentrates attention on a few previously generated summary tokens and neglects image tokens. OPERA penalizes beam-search candidates that exhibit this aggregation pattern and uses retrospection-allocation to roll back and reselect tokens when the procedure detects an unreliable summary token. HALC (Chen et al., 2024a) combines a local auto-focal grounding mechanism that corrects hallucinated tokens on the fly with a global specialized beam-search algorithm. These decoding procedures require no retraining, but their beam-level bookkeeping and specialized search operations couple the mitigation mechanism to the decoding strategy.

Contrastive decoding methods construct an auxiliary distribution that exposes undesirable language-dominant behavior and subtract that distribution from the original prediction. Visual Contrastive Decoding (VCD) (Leng et al., 2024) obtains the auxiliary distribution from a distorted version of the image. VCD contrasts predictions from the original and distorted visual inputs, thereby suppressing candidates that remain likely after reliable visual evidence has been removed. Instruction Contrastive Decoding (ICD) (Wang et al., 2024c) applies a related principle to the instruction. ICD introduces a disturbance instruction that amplifies hallucination and contrasts the disturbed distribution with the standard instruction-conditioned prediction. VCD and ICD are model-agnostic and training-free, but both methods require additional forward computation and depend on the construction of a suitable contrastive input. OPERA, HALC, VCD and ICD improve faithfulness primarily by changing candidate search or the output distribution. These approaches provide less direct control over how different attention heads accumulate visual evidence and how the model preserves that evidence across successive decoding steps.

A.4 Internal-State and Attention Intervention

Another line of work restores visual grounding inside the transformer by intervening in attention maps or intermediate representations. Attention Reallocation (AttnReal) (Tu et al., 2026) attributes hallucination to excessive attention on historical output tokens. AttnReal recycles part of this attention mass and reallocates it to visual tokens, which provides a controllable trade-off between response faithfulness and general performance with little additional computation. Paying More Attention to Image (PAI) (Liu et al., 2024c) addresses text inertia, a phenomenon in which the language backbone produces similar outputs even after the visual input changes or disappears. PAI amplifies

attention assigned to image tokens and then subtracts text-only logits from multimodal logits to suppress language-driven candidates. The PAI design therefore combines attention adjustment with contrastive logit refinement. Devils (Jiang et al., 2025) analyzes visual processing through attention and identifies the middle layers as a critical region that contains visual-information enrichment and semantic-refinement stages. Devils further associates hallucinated tokens with attention heads that interact with inconsistent objects and integrates information across heads to adjust visual attention. These analyses provide direct evidence that head-level visual inconsistency contributes to object hallucination.

Hidden-state intervention methods exploit visual information that remains encoded internally even when the final prediction underuses that information. MemVR (Zou et al., 2025) treats feed-forward networks as key-value memories and reinjects visual tokens as supplementary evidence at a middle trigger layer. MemVR activates this memory-space visual retracing when predictive uncertainty indicates that the model may have lost relevant visual evidence. VISTA (Li et al., 2025b) studies token-logit rankings across transformer layers and identifies gradual visual information loss, early excitation of semantically meaningful tokens, and visually grounded candidates that remain hidden below the final decision. VISTA combines activation-space visual steering with early-layer token-logit augmentation to recover genuine visual candidates before output selection. MemVR and VISTA demonstrate that inference-time interventions can recover latent visual evidence instead of operating only on final-layer logits.

B A Faithful-Token Prediction Margin Guarantee for Evidence-Guided Asymmetric Channel Correction

We analyze the first component of VERGE, namely the asymmetric correction of intermediate feed-forward channels by simultaneously suppressing channels associated with high-magnitude register-like visual responses and enhancing channels aligned with evidence-bearing visual responses. The goal of the analysis is not merely to show that the hidden state changes after the intervention. Instead, we establish a quantitative statement: under explicit and verifiable semantic-alignment and local-smoothness conditions, the proposed channel correction increases the logit margin of a visually faithful token against every hallucinated competitor, increases a lower bound on the faithful token probability, decreases the mass assigned to unsupported alternatives, and shifts the empirical margin distribution in a direction that yields a tighter population token-selection bound. The proof is specifically formulated at the FFN channel level so that every term corresponds directly to the mechanism implemented in the method.

B.1 Setup and main theoretical result

Given the current hidden state $h \in \mathbb{R}^d$, the gated FFN activation before the down-projection is

$$a = \phi(W_{\text{gate}}h) \odot (W_{\text{up}}h) \in \mathbb{R}^{d'}, \qquad f = W_{\text{down}}a \in \mathbb{R}^d. \tag{21}$$

The visual tokens are $\{z_i\}_{i=1}^{N_v}$. Following the method, they are partitioned by token magnitude $r_i = \|z_i\|_2$ into a high-magnitude/register subset R and an evidence subset E. For $S \in \{R, E\}$, the current hidden state retrieves a state-dependent visual prototype

$$c^S = \sum_{i \in S} \pi_i^S z_i, \qquad \pi_i^S = \frac{\exp(s_i/\tau)}{\sum_{k \in S} \exp(s_k/\tau)}, \qquad s_i = \left\langle \bar{z}_i, \bar{h} \right\rangle, \tag{22}$$

where bars denote ℓ_2 normalization. Let $w_j^\top$ denote the j-th row of W_{gate}. The channel preference scores are

$$b_j^R = \left\langle \bar{w}_j, \bar{c}^R \right\rangle, \qquad b_j^E = \left\langle \bar{w}_j, \bar{c}^E \right\rangle. \tag{23}$$

Define the nonnegative evidence score and register score vectors

$$e_j = [b_j^E]_+, \qquad s_j = \text{sigmoid}(b_j^R), \qquad e = (e_1, \ldots, e_{d'})^\top, \qquad s = (s_1, \ldots, s_{d'})^\top. \tag{24}$$

The uncertainty-dependent scaling factor is

$$r(u) = (1 - \eta) + \eta u, \qquad u \in [0, 1], \quad \eta \in [0, 1], \tag{25}$$

so that $\alpha_{\text{eff}} = \alpha r(u)$ and $\beta_{\text{eff}} = \beta r(u)$. In the non-saturated regime of the clipping operator, the suppression mask satisfies

$$m_j = 1 - \beta r(u) s_j. \tag{26}$$

The clipped case will be handled after the main derivation. Therefore, the corrected activation is

$$\widetilde{a} = a \odot m + \alpha r(u) e = a + r(u) d_0, \qquad d_0 = \alpha e - \beta(s \odot a). \tag{27}$$

Eq. (27) reveals a key structural property of the method. In the unsaturated regime, uncertainty scales rather than changes the correction direction, yielding a single composite direction d_0 that combines the evidence-enhancement vector αe with the register-suppression vector $-\beta(s \odot a)$. Consequently, a rigorous analysis can separate whether the direction itself is semantically correct from how strongly it should be applied.

Let $y^\star$ be the visually faithful next token and let $y \neq y^\star$ be any competing token. Let Φ denote the complete frozen remainder of the network from the intervened FFN output to vocabulary logits, including the residual path, subsequent transformer blocks, normalization, and the language-modeling head. Then, we can define the pairwise faithful margin as below:

$$M_y(a) = [\Phi(W_{\text{down}} a)]_{y^\star} - [\Phi(W_{\text{down}} a)]_y. \tag{28}$$

The multiclass faithful margin is

$$\Delta(a) = \min_{y \neq y^\star} M_y(a) = [\Phi(W_{\text{down}} a)]_{y^\star} - \max_{y \neq y^\star} [\Phi(W_{\text{down}} a)]_y. \tag{29}$$

A hallucinated competitor can win only if $\Delta(a) \leq 0$. Therefore, a positive lower bound on $\Delta(\widetilde{a}) - \Delta(a)$ shows that the correction makes the visually faithful token more competitive against hallucinated alternatives, directly reducing the risk of selecting an unsupported object during decoding. Then, we require two assumptions. Both are local and can be empirically diagnosed from a frozen model. First, for each competitor y, the function M_y is twice differentiable along the segment $a + qr(u)d_0$ ($q \in [0, 1]$), and its Hessian in FFN-channel coordinates obeys

$$\left\| \nabla_a^2 M_y(a + qr(u)d_0) \right\|_2 \leq L_y, \qquad L_\star = \max_{y \neq y^\star} L_y < \infty. \tag{30}$$

Second, the two prototype-induced channel directions are semantically calibrated with the faithful decision. Let $q_y = \nabla_a M_y(a)$ be the channel-space gradient of the faithful margin. For every relevant competitor y, we impose the following semantic-alignment conditions:

$$\langle q_y, e \rangle \geq \kappa_E \|e\|_2^2 - \varepsilon_E, \qquad \langle q_y, -s \odot a \rangle \geq \kappa_R \|s \odot a\|_2^2 - \varepsilon_R, \tag{31}$$

where $\kappa_E, \kappa_R > 0$ are alignment coefficients, and $\varepsilon_E, \varepsilon_R \geq 0$ denote the corresponding calibration errors. The first inequality states that channels selected by the evidence prototype are, on average, positively aligned with the direction that raises the faithful token relative to the competitor. The second inequality states that the activation carried by register-sensitive channels is, on average, anti-aligned with the faithful margin, indicating that suppressing such activation moves the decision in the desired direction. Crucially, these are not assumptions that every selected channel is individually perfect. They permit misidentified channels through the aggregate error terms ε_E and ε_R.

Under Eqs. (30)–(103), we define

$$A_y = \langle q_y, d_0 \rangle, \qquad A_\star = \min_{y \neq y^\star} A_y, \qquad D = \|d_0\|_2. \tag{32}$$

Then, we can have

$$A_\star \geq \alpha \kappa_E \|e\|_2^2 + \beta \kappa_R \|s \odot a\|_2^2 - \alpha \varepsilon_E - \beta \varepsilon_R. \tag{33}$$

For every competitor $y \neq y^\star$, the corrected pairwise margin obeys

$$M_y(\widetilde{a}) - M_y(a) \geq r(u) A_y - \frac{L_y}{2} r(u)^2 D^2, \tag{34}$$

and the multiclass margin obeys

$$\Delta(\widetilde{a}) \geq \Delta(a) + \xi(u), \qquad \xi(u) = r(u) A_\star - \frac{L_\star}{2} r(u)^2 D^2. \tag{35}$$

Hence, whenever $A_\star > 0$ and

$$0 < r(u) < \frac{2 A_\star}{L_\star D^2}, \tag{36}$$

we have $\xi(u) > 0$ and every faithful-versus-competitor margin is improved. Moreover, in the conservative regime

$$0 < r(u) \leq \frac{A_\star}{L_\star D^2}, \tag{37}$$

$\xi(u)$ is nondecreasing in uncertainty u. Therefore the uncertainty-aware scaling applies a stronger guaranteed correction exactly in the high-uncertainty regime, while remaining inside a local stability region. If $\xi(u) > 0$, the corrected faithful-token probability $\widetilde{p}_\star$ satisfies

$$\widetilde{p}_\star \geq \frac{p_\star}{p_\star + e^{-\xi(u)}(1 - p_\star)} > p_\star, \tag{38}$$

where $p_\star$ is the uncorrected softmax probability of $y^\star$. Equivalently, the total probability mass assigned to all competing tokens is strictly reduced.

Moreover, the pointwise guarantee extends to the distribution level. Let x denote a decoding context, and let $\Delta(x)$ and $\widetilde{\Delta}(x)$ denote the corresponding uncorrected and corrected multiclass faithful margins. For any target gain $\xi_0 > 0$, we define the good-correction event as

$$\mathcal{G}_{\xi_0} := \left\{ x : A_\star(x) > 0,\ 0 < r(u(x)) < \frac{2A_\star(x)}{L_\star(x) D(x)^2},\ \xi(x) \geq \xi_0 \right\}, \tag{39}$$

and let

$$\epsilon_{\mathrm{cal}}(\xi_0) = \mathbb{P}_x \left[x \notin \mathcal{G}_{\xi_0} \right] \tag{40}$$

denote the probability that the correction guarantee fails. Then, for any margin threshold $\gamma > 0$, we can have

$$\mathbb{P}[\widetilde{\Delta} \leq \gamma] \leq \mathbb{P}[\Delta \leq \gamma - \xi_0] + \epsilon_{\mathrm{cal}}(\xi_0). \tag{41}$$

Consequently, the corrected predictor has strictly less low-margin probability at threshold γ whenever

$$\mathbb{P}[\gamma - \xi_0 < \Delta \leq \gamma] > \epsilon_{\mathrm{cal}}(\xi_0). \tag{42}$$

Furthermore, suppose that the admissible inference-time configurations form a finite family $\mathcal{H}$ and that the induced margin functions are uniformly bounded in magnitude by B_0. Then, with probability at least $1 - \delta$ over n independent decoding contexts, the corrected predictor satisfies

$$\begin{aligned} R(s_{\mathrm{corr}}) \leq \widehat{F}_{\mathrm{base}}(\gamma - \xi_0) + \widehat{\epsilon}_{\mathrm{cal}}(\xi_0) \\ + \frac{2B_0}{\gamma} \sqrt{\frac{2 \log |\mathcal{H}|}{n}} + 3 \sqrt{\frac{\log(2/\delta)}{2n}}. \end{aligned} \tag{43}$$

Accordingly, the corrected predictor admits a strictly tighter margin-based population guarantee than the original frozen predictor whenever

$$\begin{aligned} \widehat{F}_{\mathrm{base}}(\gamma) - \widehat{F}_{\mathrm{base}}(\gamma - \xi_0) > \widehat{\epsilon}_{\mathrm{cal}}(\xi_0) + \Omega_{\mathrm{corr}}(n, \gamma, \delta) \\ - \Omega_{\mathrm{base}}(n, \gamma, \delta). \end{aligned} \tag{44}$$

B.2 Proof of Theorem 1

From Eq. (27), the channel perturbation induced by the proposed correction is

$$\delta a = \widetilde{a} - a = r(u) d_0. \tag{45}$$

To characterize how this perturbation affects the faithful margin against a fixed competitor $y \neq y^\star$, we parameterize the line segment connecting the original and corrected FFN activations by

$$\zeta(q) = a + q \delta a, \qquad q \in [0, 1]. \tag{46}$$

We then characterize the variation of the pairwise margin along the parameterized intervention trajectory by defining the scalar function

$$\varphi_y(q) = M_y(\zeta(q)) = M_y(a + q \delta a)\,. \tag{47}$$

By the chain rule, the first derivative of $\varphi_y(q)$ is

$$\varphi_y'(q) = \langle \nabla M_y(a + q \delta a), \delta a \rangle\,, \tag{48}$$

where $\nabla M_y(a + q\delta a)$ denotes the gradient of the pairwise margin M_y with respect to the FFN channel activation, evaluated at $a + q\delta a$. Differentiating once more gives

$$\varphi_y''(q) = \delta a^\top \nabla^2 M_y(a + q\delta a)\delta a. \tag{49}$$

Applying the second-order Taylor theorem with integral remainder to φ_y between $q = 0$ and $q = 1$ yields

$$\begin{aligned} M_y(a + \delta a) = M_y(a) + \langle \nabla M_y(a), \delta a \rangle \\ + \int_0^1 (1 - q)\, \delta a^\top \nabla^2 M_y(a + q\delta a)\delta a\, dq. \end{aligned} \tag{50}$$

Under the local smoothness condition in Eq. (30), the Hessian matrix of the pairwise margin function M_y with respect to the FFN channel activation is uniformly bounded along the parameterized intervention trajectory. Specifically, for every $q \in [0, 1]$, we can have

$$\left\| \nabla^2 M_y(a + q\delta a) \right\|_2 \le L_y, \tag{51}$$

where $\nabla^2 M_y(a + q\delta a)$ denotes the Hessian matrix of M_y evaluated at $a + q\delta a$. This matrix collects the second-order partial derivatives of the pairwise margin with respect to the FFN channel coordinates and characterizes its local curvature along the intervention trajectory. Accordingly, the quadratic form $\delta a^\top \nabla^2 M_y(a + q\delta a)\delta a$ measures the second-order variation of the pairwise margin along the perturbation direction δa. Since the Hessian matrix is symmetric, its quadratic form satisfies

$$\delta a^\top \nabla^2 M_y(a + q\delta a)\delta a \ge - \left\| \nabla^2 M_y(a + q\delta a) \right\|_2 \|\delta a\|_2^2 \ge -L_y \|\delta a\|_2^2 . \tag{52}$$

Substituting Eq. (52) into the Taylor expansion in Eq. (50), and using $\delta a = r(u)d_0$, we can have

$$M_y(\widetilde{a}) - M_y(a) \ge \langle \nabla M_y(a), r(u)d_0 \rangle - \int_0^1 (1 - q) L_y r(u)^2 \|d_0\|_2^2 \, dq. \tag{53}$$

Recalling the definitions

$$q_y = \nabla M_y(a), \qquad A_y = \langle q_y, d_0 \rangle , \qquad D = \|d_0\|_2 , \tag{54}$$

the preceding inequality can be written as

$$\begin{aligned} M_y(\widetilde{a}) - M_y(a) &\ge r(u)A_y - L_y r(u)^2 D^2 \int_0^1 (1 - q)\, dq \\ &= r(u)A_y - \frac{L_y}{2} r(u)^2 D^2. \end{aligned} \tag{55}$$

Next, we decompose the first-order gain into the two mechanisms implemented by the method. By definition, we can get

$$A_y = \langle q_y, \alpha e - \beta(s \odot a) \rangle \tag{56}$$

$$= \alpha \langle q_y, e \rangle + \beta \langle q_y, -s \odot a \rangle . \tag{57}$$

Applying Eq. (103), we can have

$$A_y \ge \alpha\big(\kappa_E \|e\|_2^2 - \varepsilon_E\big) + \beta\big(\kappa_R \|s \odot a\|_2^2 - \varepsilon_R\big) \tag{58}$$

$$= \alpha\kappa_E \|e\|_2^2 + \beta\kappa_R \|s \odot a\|_2^2 - \alpha\varepsilon_E - \beta\varepsilon_R. \tag{59}$$

Taking the minimum over $y \neq y^\star$ yields Eq. (33). This expression gives a direct quantitative interpretation of the asymmetric design. The enhancement term contributes in proportion to the squared energy of channels recognized by the evidence prototype, while the suppression term contributes in proportion to the squared activated energy of channels recognized by the register prototype. The errors ε_E and ε_R quantify the price paid for imperfect channel identification. Thus, suppression and enhancement are not redundant. Either term can compensate for a weak contribution from the other, and both enter the first-order margin gain additively.

The second-order curvature penalty is governed by the perturbation energy D^2, whose structure further reveals the benefit of separating the evidence-enhancement and register-suppression directions. To quantify the interaction between these two components, we define

$$E_c = \|e\|_2 , \qquad R_c = \|s \odot a\|_2 , \qquad \chi = \frac{|\langle e, s \odot a \rangle|}{E_c R_c} \in [0, 1]. \tag{60}$$

When $E_c R_c = 0$, we set $\chi = 0$ by convention. Here, χ measures the directional coherence between the evidence-enhancement vector e and the activated register-sensitive component $s \odot a$ in the FFN channel space. Recalling that $d_0 = \alpha e - \beta(s \odot a)$, the perturbation energy can be expanded as

$$D^2 = \|\alpha e - \beta(s \odot a)\|_2^2 \tag{61}$$

$$= \alpha^2 E_c^2 + \beta^2 R_c^2 - 2\alpha\beta \langle e, s \odot a \rangle \tag{62}$$

$$\leq \alpha^2 E_c^2 + \beta^2 R_c^2 + 2\alpha\beta\chi E_c R_c. \tag{63}$$

The bound shows that the interaction term in the perturbation energy is directly controlled by the coherence χ. A smaller χ indicates weaker directional overlap between the channels emphasized by evidence enhancement and those targeted by register suppression, thereby limiting the additional perturbation energy induced by their joint intervention. This separation is particularly favorable since the two correction components contribute additively to the first-order faithful-margin gain, whereas their interaction enters the second-order curvature penalty through D^2. Consequently, low channel-space coherence limits the second-order interaction cost while leaving the additive first-order semantic gain unaffected. This provides a principled justification for constructing separate state-conditioned evidence and register prototypes rather than applying a single global rescaling to the FFN channels.

The clipping operator does not alter the preceding pointwise margin argument. In the general case, define the realized suppression amount as

$$\rho_j(u) = 1 - m_j = 1 - \mathrm{clip}(1 - \beta r(u) s_j, c_{\min}, 1), \qquad 0 \leq \rho_j(u) \leq 1 - c_{\min}.$$

Let $\rho(u) = (\rho_1(u), \ldots, \rho_{d'}(u))^\top$. The actual perturbation then becomes

$$\delta a = \widetilde{a} - a = \alpha r(u) e - \rho(u) \odot a.$$

Accordingly, the same Taylor argument applies after replacing $\beta r(u)(s \odot a)$ with $\rho(u) \odot a$, provided that the realized suppression direction $-\rho(u) \odot a$ satisfies the corresponding semantic-alignment condition. Clipping therefore preserves the pointwise margin-improvement argument while bounding the suppression magnitude by $1 - c_{\min}$, preventing excessive attenuation of FFN activity. The closed-form positivity and monotonicity conditions in Eqs. (36)–(37) characterize the unclipped regime; once clipping is active, the dependence on uncertainty becomes piecewise, while the subsequent margin and probability guarantees remain valid whenever the realized margin gain is positive.

We now lift the pairwise result to the multiclass margin. Eq. (34) holds for every $y \neq y^\star$. Since $A_y \geq A_\star$ and $L_y \leq L_\star$, we can have

$$M_y(\widetilde{a}) \geq M_y(a) + r(u) A_\star - \frac{L_\star}{2} r(u)^2 D^2 = M_y(a) + \xi(u). \tag{64}$$

Taking the minimum over all competitors gives

$$\Delta(\widetilde{a}) = \min_{y \neq y^\star} M_y(\widetilde{a}) \tag{65}$$

$$\geq \min_{y \neq y^\star} \big(M_y(a) + \xi(u)\big) \tag{66}$$

$$= \Delta(a) + \xi(u), \tag{67}$$

which proves Eq. (35). The function $\xi(u)$ is a concave quadratic in $r(u)$. If $A_\star > 0$, then $\xi(u) > 0$ whenever

$$0 < r(u) < \frac{2A_\star}{L_\star D^2}, \tag{68}$$

which proves Eq. (36). Moreover, since $\frac{d\xi}{dr} = A_\star - L_\star D^2 r$, we can have $d\xi/dr \geq 0$ for $r \leq A_\star/(L_\star D^2)$. Since $dr/du = \eta \geq 0$, the chain rule gives

$$\frac{d\xi}{du} = \eta \left(A_\star - L_\star D^2 r(u)\right) \geq 0 \tag{69}$$

throughout the range specified in Eq. (37). Hence, a more uncertain decoding state induces a stronger intervention through $r(u)$, and the resulting increase in correction strength leads to a larger guaranteed margin improvement. The restriction on $r(u)$ ensures that this monotonic behavior holds only in the regime where the first-order semantic gain remains dominant over the second-order curvature

penalty. It remains to convert the margin improvement into a probability statement. For the uncorrected logits, the faithful softmax probability can be written entirely in terms of pairwise margins:

$$p_\star = \frac{\exp(z_{y^\star})}{\exp(z_{y^\star}) + \sum_{y \neq y^\star} \exp(z_y)} = \frac{1}{1 + \sum_{y \neq y^\star} \exp[-M_y(a)]}, \tag{70}$$

where $z_y = [\Phi(W_{\text{down}} a)]_y$ denotes the uncorrected vocabulary logit associated with token y, and $z_{y^\star}$ denotes the corresponding logit of the visually faithful token $y^\star$. From Eq. (64), we can obtain

$$\exp[-M_y(\widetilde{a})] \leq \exp[-\xi(u)] \exp[-M_y(a)]. \tag{71}$$

Therefore, we can have

$$\widetilde{p}_\star = \frac{1}{1 + \sum_{y \neq y^\star} \exp[-M_y(\widetilde{a})]} \tag{72}$$

$$\geq \frac{1}{1 + e^{-\xi(u)} \sum_{y \neq y^\star} \exp[-M_y(a)]}. \tag{73}$$

Using Eq. (70), we can get

$$\sum_{y \neq y^\star} \exp[-M_y(a)] = \frac{1 - p_\star}{p_\star}. \tag{74}$$

Substitution Eq. (74) into Eq. (73) yields

$$\widetilde{p}_\star \geq \frac{p_\star}{p_\star + e^{-\xi(u)}(1 - p_\star)}, \tag{75}$$

which establishes Eq. (38). The condition $\xi(u) > 0$ ensures that $e^{-\xi(u)} < 1$, which strictly reduces the aggregate contribution of competing tokens in the softmax denominator. Consequently, the resulting lower bound on the corrected faithful-token probability is strictly larger than the original probability, yielding $\widetilde{p}_\star > p_\star$. Equivalently, the total probability mass assigned to all competing tokens is strictly decreased after correction. This completes the proof.

B.3 Distribution and Generalization Consequence

Theorem 1 is pointwise, i.e., it explains what happens for a particular decoding context when the register/evidence prototypes successfully identify a beneficial correction direction. We now translate the pointwise guarantee into a distribution-level statement suitable for hallucination-risk analysis. Let x denote a decoding context, including the image, prompt, and previous generated tokens. Let $\Delta(x)$ and $\widetilde{\Delta}(x)$ denote the uncorrected and corrected multiclass faithful margins. Since Theorem 1 gives $\widetilde{\Delta}(x) \geq \Delta(x) + \xi.$ for any threshold $\gamma > 0$, we can have

$$\{\widetilde{\Delta}(x) \leq \gamma\} \subseteq \{\Delta(x) \leq \gamma - \xi\} \cup \{x \notin \mathcal{G}_{\xi_0}\}. \tag{76}$$

Taking probabilities yields

$$\mathbb{P}[\widetilde{\Delta} \leq \gamma] \leq \mathbb{P}[\Delta \leq \gamma - \xi_0] + \epsilon_{\text{cal}}(\xi_0). \tag{77}$$

Eq. (77) provides an upper bound on the low-margin probability of the corrected predictor, but the bound alone does not imply a strict improvement over the baseline predictor. To establish a strict reduction at threshold γ, the upper bound on the right-hand side of Eq. (77) must itself be smaller than the baseline low-margin probability $\mathbb{P}[\Delta \leq \gamma]$. A sufficient condition is therefore

$$\mathbb{P}[\Delta \leq \gamma - \xi_0] + \epsilon_{\text{cal}}(\xi_0) < \mathbb{P}[\Delta \leq \gamma]. \tag{78}$$

Under this condition, Eq. (77) immediately gives

$$\mathbb{P}[\widetilde{\Delta} \leq \gamma] < \mathbb{P}[\Delta \leq \gamma], \tag{79}$$

which establishes a strict reduction in low-margin probability after correction. Rearranging Eq. (78) gives the equivalent sufficient condition

$$\mathbb{P}[\Delta \leq \gamma] - \mathbb{P}[\Delta \leq \gamma - \xi_0] > \epsilon_{\text{cal}}(\xi_0). \tag{80}$$

Since $\xi_0 > 0$, the event $\{\Delta \le \gamma - \xi_0\}$ is contained in $\{\Delta \le \gamma\}$. Therefore, the difference on the left-hand side of Eq. (80) can be expressed as

$$\mathbb{P}[\Delta \le \gamma] - \mathbb{P}[\Delta \le \gamma - \xi_0] = \mathbb{P}[\gamma - \xi_0 < \Delta \le \gamma]. \tag{81}$$

The probability in Eq. (81) measures the baseline probability mass contained in the recoverable interval $(\gamma - \xi_0, \gamma]$. For every decoding context in this interval, a guaranteed margin increase of at least ξ_0 moves the corrected margin above the threshold γ. Thus, Eq. (80) shows that strict distribution improvement is guaranteed whenever the recoverable near-boundary mass exceeds the probability that the correction guarantee fails.

This distribution improvement can be further related to the population token-selection risk through a standard margin argument. Let s denote a vocabulary scoring function, and let

$$\Delta_s(x, y^\star) = s_{y^\star}(x) - \max_{y \ne y^\star} s_y(x) \tag{82}$$

denote its multiclass faithful margin for decoding context x and visually faithful token $y^\star$. The corresponding population token-selection risk is defined as

$$R(s) = \mathbb{P}[\Delta_s(x, y^\star) \le 0]. \tag{83}$$

The event $\Delta_s(x, y^\star) \le 0$ indicates that at least one competing token receives a score no smaller than the score assigned to $y^\star$. Therefore, the quantity $R(s)$ measures the probability that the scoring function fails to rank the faithful token above all competing alternatives. For any margin threshold $\gamma > 0$, we define the ramp loss as below:

$$\psi_\gamma(r) = \begin{cases} 1, & r \le 0, \\ 1 - r/\gamma, & 0 < r < \gamma, \\ 0, & r \ge \gamma. \end{cases} \tag{84}$$

The ramp loss satisfies

$$\mathbf{1}[r \le 0] \le \psi_\gamma(r) \le \mathbf{1}[r \le \gamma], \tag{85}$$

and its Lipschitz constant is $1/\gamma$. The first inequality in Eq. (85) yields

$$\begin{aligned} R(s) &= \mathbb{E}[\mathbf{1}[\Delta_s(x, y^\star) \le 0]] \\ &\le \mathbb{E}[\psi_\gamma(\Delta_s(x, y^\star))]. \end{aligned}$$

Let $\mathcal{S}$ denote a class of vocabulary scoring functions, and let

$$\mathcal{M}_\mathcal{S} = \{(x, y^\star) \mapsto \Delta_s(x, y^\star) : s \in \mathcal{S}\} \tag{86}$$

denote the corresponding class of faithful-margin functions. A standard Rademacher generalization bound can be applied to the ramp-loss class to control its population expectation by the corresponding empirical quantity. With probability at least $1 - \delta$ over n independent decoding contexts, the following inequality holds:

$$\begin{aligned} \mathbb{E}[\psi_\gamma(\Delta_s(x, y^\star))] \le \frac{1}{n} \sum_{i=1}^{n} \psi_\gamma(\Delta_s(x_i, y_i^\star)) \\ + 2\widehat{\mathfrak{R}}_n\left(\psi_\gamma \circ \mathcal{M}_\mathcal{S}\right) + 3\sqrt{\frac{\log(2/\delta)}{2n}}. \end{aligned} \tag{87}$$

Since ψ_γ is $1/\gamma$-Lipschitz, the Rademacher contraction inequality gives

$$\widehat{\mathfrak{R}}_n\left(\psi_\gamma \circ \mathcal{M}_\mathcal{S}\right) \le \frac{1}{\gamma} \widehat{\mathfrak{R}}_n(\mathcal{M}_\mathcal{S}). \tag{88}$$

The second inequality in Eq. (85) also gives

$$\frac{1}{n} \sum_{i=1}^{n} \psi_\gamma(\Delta_s(x_i, y_i^\star)) \le \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[\Delta_s(x_i, y_i^\star) \le \gamma]. \tag{89}$$

Then, we define the empirical low-margin frequency as

$$\widehat{F}_s(\gamma) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}[\Delta_s(x_i, y_i^\star) \le \gamma]. \tag{90}$$

Combining Eqs. (86)–(90) yields

$$R(s) \le \widehat{F}_s(\gamma) + \frac{2}{\gamma}\widehat{\mathfrak{R}}_n(\mathcal{M}_{\mathcal{S}}) + 3\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{91}$$

Eq. (91) shows that the population token-selection risk is controlled by the empirical frequency of low-margin predictions and the statistical complexity of the induced margin class.

VERGE does not update the backbone parameters, and the corrected predictor therefore varies only through its inference-time configuration. Let $\mathcal{H}$ denote the finite family of admissible inference-time hyperparameter configurations, and let $\mathcal{M}_{\mathcal{H}}$ denote the corresponding family of induced margin functions. Assume that every margin function in $\mathcal{M}_{\mathcal{H}}$ is uniformly bounded in magnitude by B_0. Massart's finite-class lemma then provides the following upper bound on the empirical Rademacher complexity:

$$\widehat{\mathfrak{R}}_n(\mathcal{M}_{\mathcal{H}}) \le B_0\sqrt{\frac{2\log|\mathcal{H}|}{n}}. \tag{92}$$

Eq. (92) shows that the additional statistical complexity associated with inference-time configuration selection depends only logarithmically on the size of $\mathcal{H}$.

The pointwise event inclusion underlying Eq. (77) also yields an empirical counterpart. For every sampled decoding context x_i, we can have

$$\mathbf{1}[\widetilde{\Delta}(x_i) \le \gamma] \le \mathbf{1}[\Delta(x_i) \le \gamma - \xi_0] + \mathbf{1}[x_i \notin \mathcal{G}_{\xi_0}]. \tag{93}$$

Averaging Eq. (93) over all n decoding contexts gives

$$\widehat{F}_{s_{\mathrm{corr}}}(\gamma) \le \widehat{F}_{s_{\mathrm{base}}}(\gamma - \xi_0) + \widehat{\epsilon}_{\mathrm{cal}}(\xi_0), \tag{94}$$

where

$$\widehat{\epsilon}_{\mathrm{cal}}(\xi_0) = \frac{1}{n}\sum_{i=1}^{n}\mathbf{1}[x_i \notin \mathcal{G}_{\xi_0}] \tag{95}$$

denotes the empirical probability that the correction guarantee fails to provide a margin increase of at least ξ_0. Applying Eq. (91) to the corrected predictor and substituting Eqs. (92) and (94) yields

$$\begin{aligned} R(s_{\mathrm{corr}}) \le{}& \widehat{F}_{s_{\mathrm{base}}}(\gamma - \xi_0) + \widehat{\epsilon}_{\mathrm{cal}}(\xi_0) \\ &+ \frac{2B_0}{\gamma}\sqrt{\frac{2\log|\mathcal{H}|}{n}} + 3\sqrt{\frac{\log(2/\delta)}{2n}}. \end{aligned} \tag{96}$$

For compactness, we define

$$\Omega_{s_{\mathrm{corr}}}(n,\gamma,\delta) = \frac{2B_0}{\gamma}\sqrt{\frac{2\log|\mathcal{H}|}{n}} + 3\sqrt{\frac{\log(2/\delta)}{2n}}. \tag{97}$$

The corrected risk bound can then be written as

$$R(s_{\mathrm{corr}}) \le \widehat{F}_{s_{\mathrm{base}}}(\gamma - \xi_0) + \widehat{\epsilon}_{\mathrm{cal}}(\xi_0) + \Omega_{s_{\mathrm{corr}}}(n,\gamma,\delta). \tag{98}$$

The same margin argument applied to the original frozen decoder gives

$$R(s_{\mathrm{base}}) \le \widehat{F}_{s_{\mathrm{base}}}(\gamma) + \Omega_{s_{\mathrm{base}}}(n,\gamma,\delta), \tag{99}$$

where $\Omega_{s_{\mathrm{base}}}(n,\gamma,\delta)$ collects the corresponding generalization terms for the baseline predictor. A direct comparison of Eqs. (98) and (99) shows that the corrected predictor admits a strictly tighter margin-based upper bound whenever

$$\begin{aligned} &\widehat{F}_{s_{\mathrm{base}}}(\gamma - \xi_0) + \widehat{\epsilon}_{\mathrm{cal}}(\xi_0) + \Omega_{s_{\mathrm{corr}}}(n,\gamma,\delta) \\ &\quad < \widehat{F}_{s_{\mathrm{base}}}(\gamma) + \Omega_{s_{\mathrm{base}}}(n,\gamma,\delta). \end{aligned} \tag{100}$$

Rearranging Eq. (100) gives the equivalent sufficient condition

$$\begin{aligned} \widehat{F}_{s_{\mathrm{base}}}(\gamma) - \widehat{F}_{s_{\mathrm{base}}}(\gamma - \xi_0) > \widehat{\epsilon}_{\mathrm{cal}}(\xi_0) + \Omega_{s_{\mathrm{corr}}}(n,\gamma,\delta) \\ - \Omega_{s_{\mathrm{base}}}(n,\gamma,\delta). \end{aligned} \tag{101}$$

The left-hand side of Eq. (101) measures the empirical mass of baseline contexts whose faithful margins lie in the recoverable interval $(\gamma - \xi_0, \gamma]$. The right-hand side accounts for the empirical correction-failure probability together with the difference between the corrected and baseline complexity terms. Since the proposed intervention introduces no additional trainable backbone parameters and only selects from a finite inference-time configuration family, Eq. (92) bounds the additional configuration-dependent complexity by $O(\sqrt{\log |\mathcal{H}|/n})$. The comparison in Eq. (101) therefore shows that the corrected predictor obtains a tighter margin-based guarantee when the recoverable low-margin mass is sufficiently large to dominate both the correction-failure probability and the additional inference-time configuration complexity.

Interpretation for the method design. The preceding analysis provides a direct theoretical justification for the design of the first module. The register and evidence prototypes construct two input-dependent channel-space directions that play complementary roles in the correction. Under the semantic-alignment conditions, the evidence-enhancement direction is positively aligned with an increase in the faithful-token margin, while the register-suppression direction removes activation that is negatively aligned with that margin. Their asymmetric combination therefore produces an additive first-order semantic gain. The local smoothness condition controls the nonlinear distortion introduced by the frozen downstream network and yields an explicit range of intervention strengths within which this first-order gain dominates the second-order curvature penalty. Within the corresponding conservative regime, the uncertainty-dependent scaling factor increases the guaranteed margin improvement as predictive uncertainty rises, which provides a principled basis for applying stronger correction to less reliable decoding states. The use of separate register and evidence prototypes further supports this mechanism since low channel-space coherence χ reduces the interaction-induced contribution to the perturbation energy in Eq. (63), while the first-order gains contributed by the two semantic directions remain additive and are governed independently by their alignment conditions. Under these conditions, the resulting margin improvement strengthens the preference for the visually faithful token and increases its softmax probability according to Eq. (38). At the distribution level, the same improvement shifts probability mass away from the low-margin region according to Eq. (77), indicating that contexts with margins close to the decision threshold can be moved into a more reliable regime. The generalization analysis further shows that, under the condition in Eq. (101), this distribution improvement can be translated into a tighter population-level token-selection guarantee. Together, these results connect the channel-wise correction to improved faithful-token preference, a more favorable margin distribution, and stronger population-level reliability. The theoretical analysis therefore explains why evidence enhancement, register suppression, uncertainty-aware modulation, and bounded intervention strength are complementary components of the proposed channel-wise correction mechanism.

C An Information-Sensitive Generalization Guarantee for Uncertainty-Triggered Visual Retracing

This section develops a unified theoretical characterization of the visual retracing mechanism in VERGE. Rather than viewing visual retracing as an unconstrained feature injection, we formulate it as an uncertainty-triggered information-recovery operator that selectively reintroduces image-specific evidence when the current representation exhibits weakened visual grounding. The analysis then establishes how the recovered target-relevant information can be translated into a population-level generalization advantage.

The proof adopts a representation-centric generalization analysis. It first upper-bounds the original prediction loss by a Lipschitz surrogate defined on the internal representation. It then constructs a high-probability class-conditional typical set from conditional self-information and decomposes the gap between population risk and empirical risk into two sources: the probability mass assigned to atypical representations and the empirical-frequency deviations of typical representation cells. Concentration inequalities are then applied to control these deviations, which yields a representation complexity term determined by conditional entropy. The resulting entropy term is subsequently expressed through mutual information, making explicit how target-relevant information contributes to a tighter generalization guarantee. As a result, the final bound characterizes the actual visual retracing mechanism rather than an idealized form of information injection.

C.1 SETUP AND MAIN THEORETICAL RESULT

We consider a decoding context drawn from the conditional distribution in which the uncertainty trigger is activated. We use $T \in \{0, 1\}$ to denote the trigger indicator, and all probabilities in the main derivation are conditioned on $T = 1$. The variable X collects the image, prompt, and previously generated tokens that define the current decoding context. The variable $Y \in \{1, \ldots, K\}$ denotes the visually faithful next-token label. We use $Z = [z_1, \ldots, z_{N_v}]^\top \in \mathbb{R}^{N_v \times d}$ to denote the original projected visual tokens. We use $h \in \mathbb{R}^d$ to denote the hidden state at the first triggered layer, and we use $\widetilde{f} \in \mathbb{R}^d$ to denote the feed-forward output after the preceding channel-wise correction.

The visual retracing operator constructs a temporary key–value memory $K_v = s_1 Z$ and $V_v = s_2 Z$. The raw retrieved representation is

$$R_0 = (hK_v^\top)V_v = s_1 s_2 \sum_{j=1}^{N_v} \langle h, z_j \rangle z_j. \tag{102}$$

The method subsequently aligns the magnitude of the retrieved representation with the current feed-forward pathway. The aligned retrieval is

$$R = a(h, \widetilde{f}, Z) R_0, \qquad a(h, \widetilde{f}, Z) = \frac{d^{-1}\|\widetilde{f}\|_1}{d^{-1}\|R_0\|_1 + \epsilon_0}, \tag{103}$$

The retraced feed-forward state is given by the following expression.

$$\widehat{f} = (1 - \lambda)\widetilde{f} + \lambda R, \qquad \lambda \in [0, 1]. \tag{104}$$

We denote the representation before visual retracing by S_0 and the representation after visual retracing by S_1. The subsequent analysis treats these two representation states directly. The exact target-information gain induced by visual retracing is defined as

$$G_{\mathrm{VR}} = I(S_1; Y \mid T = 1) - I(S_0; Y \mid T = 1). \tag{105}$$

The analysis uses a bounded loss ℓ and a frozen suffix map ψ that maps an intermediate representation to the scalar score associated with the prediction under analysis. The loss satisfies $0 \le \ell \le M_\ell$ and is L_ℓ-Lipschitz in the score, while ψ is L_ψ-Lipschitz on a common activation envelope containing both S_0 and S_1. For each target class y_c, the representation is parameterized as $S = \phi_{y_c}(V_1, \ldots, V_m)$, where $V_1, \ldots, V_m$ denote nuisance factors that may alter the representation while leaving the target class unchanged. We define the class-conditional self-information as

$$q_{y_c}(s) = -\log p(s \mid Y = y_c, T = 1), \tag{106}$$

which measures how unlikely a representation state s is under the conditional distribution associated with y_c. We further assume that $q_{y_c}(S)$ has bounded coordinate-wise sensitivity. Specifically, for any nuisance coordinate V_j, replacing v_j by an alternative value v_j' while keeping all remaining coordinates fixed satisfies

$$\left| q_{y_c}(\phi_{y_c}(v_1, \ldots, v_j, \ldots, v_m)) - q_{y_c}\big(\phi_{y_c}(v_1, \ldots, v_j', \ldots, v_m)\big) \right| \le c_S. \tag{107}$$

Thus, c_S quantifies the largest change in the class-conditional self-information that can be induced by any single nuisance factor. This condition prevents an individual nuisance variation from causing an arbitrarily large change in the likelihood of the representation and provides the bounded-difference property required for the subsequent concentration analysis. For a generic representation S, we define the conditional population risk, empirical risk, and generalization gap as

$$R(S) = \mathbb{E}[\ell(\psi(S), Y) \mid T = 1], \qquad \widehat{R}_n(S) = \frac{1}{n}\sum_{i=1}^{n} \ell(\psi(S_i), Y_i), \qquad \mathrm{Gen}_n(S) = R(S) - \widehat{R}_n(S) \tag{108}$$

where $\{(S_i, Y_i)\}_{i=1}^n$ are i.i.d. samples from the triggered distribution. We define the surrogate loss as follows:

$$\widehat{\ell}_S(s, y) = \ell(\psi(0), y) + L_\ell L_\psi \|s\|_2, \tag{109}$$

and we define the empirical surrogate residual as below:

$$\rho_S = \frac{1}{n}\sum_{i=1}^{n} \Big[\widehat{\ell}_S(S_i, Y_i) - \ell(\psi(S_i), Y_i)\Big] \ge 0. \tag{110}$$

We fix the constant $\gamma \in (0,1)$ and $\delta \in (0,1)$. The bounded-difference condition in Eq. (107) allows McDiarmid's inequality to control the upper tail of the class-conditional self-information. We choose the concentration radius such that the corresponding upper-tail probability is at most $\gamma/\sqrt{n}$. At the boundary of this requirement, the exponential tail factor satisfies

$$\exp\left(-\frac{2\epsilon^2}{mc_S^2}\right) = \frac{\gamma}{\sqrt{n}}.$$

Taking logarithms on both sides gives

$$\frac{2\epsilon^2}{mc_S^2} = \log\frac{\sqrt{n}}{\gamma}.$$

The nonnegative solution of this relation yields the concentration radius

$$\epsilon_S(n,\gamma) = \epsilon = c_S\sqrt{\frac{m}{2}\log\frac{\sqrt{n}}{\gamma}}. \tag{111}$$

We use constants $A_1, A_2, A_3, A_4 > 0$ whose explicit values are stated after Eq. (163). With probability at least $1-\delta$, every admissible representation S satisfies

$$\begin{aligned}\mathrm{Gen}_n(S) \le \rho_S + A_1\frac{\gamma}{\sqrt{n}} + A_2\frac{\sqrt{\gamma\log(4K/\delta)}}{n^{3/4}} + A_4\frac{\log(4K/\delta)}{n} \\ + A_3\sqrt{\frac{H(S\mid Y,T=1)+\Gamma_S}{n}},\end{aligned} \tag{112}$$

The auxiliary quantity Γ_S is specified by the following expression:

$$\Gamma_S = \epsilon_S + \log\frac{2K}{\delta}. \tag{113}$$

The conditional entropy in Eq. (112) admits the exact information decomposition

$$H(S\mid Y,T=1) = H(S\mid T=1) - I(S;Y\mid T=1). \tag{114}$$

Consequently, we define

$$V_j = H(S_j\mid T=1) - I(S_j;Y\mid T=1) + \Gamma_{S_j}, \qquad j\in\{0,1\}. \tag{115}$$

If $V_1 < V_0$, then the representation-dependent square-root term in the generalization bound is strictly smaller after visual retracing. More explicitly, we define

$$\Delta H = H(S_1\mid T=1) - H(S_0\mid T=1), \qquad \Delta\Gamma = \Gamma_{S_1} - \Gamma_{S_0}. \tag{116}$$

The relation $I(S_1;Y\mid T=1) - I(S_0;Y\mid T=1) = G_{\mathrm{VR}}$ implies

$$V_0 - V_1 = G_{\mathrm{VR}} - \Delta H - \Delta\Gamma. \tag{117}$$

The representation-complexity term is strictly tightened under the following sufficient condition

$$G_{\mathrm{VR}} > \Delta H + \Delta\Gamma, \tag{118}$$

This condition guarantees a strict reduction of the representation-dependent complexity term. The complete upper bound is strictly tightened whenever the reduction

$$\frac{A_3}{\sqrt{n}}\frac{G_{\mathrm{VR}} - \Delta H - \Delta\Gamma}{\sqrt{V_0}+\sqrt{V_1}} \tag{119}$$

exceeds the retracing-induced increase $\rho_{S_1} - \rho_{S_0}$ in the surrogate residual.

C.2 Derivation of the generalization guarantee

For every representation state s and target y, the Lipschitz properties of ℓ and ψ imply the following bound

$$\begin{aligned}\ell(\psi(s), y) &\le \ell(\psi(0), y) + L_\ell |\psi(s) - \psi(0)| \\ &\le \ell(\psi(0), y) + L_\ell L_\psi \|s\|_2 = \widehat{\ell}_S(s, y).\end{aligned} \tag{120}$$

The relation in Eq. (120) yields an exact decomposition of the generalization gap:

$$\begin{aligned}\mathrm{Gen}_n(S) = &\Big(R(S) - \mathbb{E}[\widehat{\ell}_S(S, Y) \mid T = 1]\Big) \\ &+ \Big(\mathbb{E}[\widehat{\ell}_S(S, Y) \mid T = 1] - \frac{1}{n}\sum_{i=1}^{n} \widehat{\ell}_S(S_i, Y_i)\Big) \\ &+ \frac{1}{n}\sum_{i=1}^{n}\Big(\widehat{\ell}_S(S_i, Y_i) - \ell(\psi(S_i), Y_i)\Big).\end{aligned} \tag{121}$$

The first bracket in Eq. (121) is non-positive because the surrogate dominates the original loss pointwise. The last term equals ρ_S by definition. Hence, the remaining analysis only needs to control the discrepancy between the surrogate population risk and its empirical counterpart:

$$\mathrm{Gen}_n(S) \le \mathbb{E}[\widehat{\ell}_S(S, Y) \mid T = 1] - \frac{1}{n}\sum_{i=1}^{n} \widehat{\ell}_S(S_i, Y_i) + \rho_S. \tag{122}$$

The next part of the analysis converts the representation distribution into a finite set of high-probability cells. Under the conditional law $p(s \mid y_c, T = 1)$, we define the random self-information as follows:

$$Q_c = -\log p(S \mid Y = y_c, T = 1). \tag{123}$$

Its conditional expectation is exactly the class-conditional entropy:

$$\begin{aligned}\mathbb{E}[Q_c \mid Y = y_c, T = 1] &= -\sum_s p(s \mid y_c, T = 1) \log p(s \mid y_c, T = 1) \\ &= H(S \mid Y = y_c, T = 1).\end{aligned} \tag{124}$$

Since Q_c changes by at most c_S when one of the m nuisance coordinates is replaced, McDiarmid's inequality gives the following tail bound

$$\mathbb{P}(Q_c - H(S \mid y_c, T = 1) \ge \epsilon \mid y_c, T = 1) \le \exp\left(-\frac{2\epsilon^2}{m c_S^2}\right). \tag{125}$$

Substituting the concentration radius derived in Eq. (111) into Eq. (125) gives the following class-uniform probability estimate

$$\mathbb{P}(Q_c > H(S \mid y_c, T = 1) + \epsilon_S \mid y_c, T = 1) \le \frac{\gamma}{\sqrt{n}}. \tag{126}$$

This class-uniform estimate motivates the following typical representation set

$$\mathcal{T}_{c,\gamma} = \{s : -\log p(s \mid y_c, T = 1) \le H(S \mid y_c, T = 1) + \epsilon_S\}. \tag{127}$$

Every $s \in \mathcal{T}_{c,\gamma}$ satisfies

$$-\log p(s \mid y_c, T = 1) \le H(S \mid y_c, T = 1) + \epsilon_S. \tag{128}$$

Equivalently, each typical representation state carries conditional probability at least

$$p(s \mid y_c, T = 1) \ge \exp[-H(S \mid y_c, T = 1) - \epsilon_S]. \tag{129}$$

Let $T_c = |\mathcal{T}_{c,\gamma}|$ denote the number of representation states in the typical set. Since $\mathcal{T}_{c,\gamma}$ is a subset of the complete representation space, its total conditional probability satisfies

$$\begin{aligned}1 \ge &\sum_{s \in \mathcal{T}_{c,\gamma}} p(s \mid y_c, T = 1) \\ &\ge T_c \exp[-H(S \mid y_c, T = 1) - \epsilon_S].\end{aligned} \tag{130}$$

Rearranging Eq. (130) gives

$$T_c = |\mathcal{T}_{c,\gamma}| \le \exp[H(S \mid y_c, T=1) + \epsilon_S]\,. \tag{131}$$

We next decompose the discrepancy between the population and empirical surrogate risks over the class-conditional representation cells. We enumerate the typical cells for class y_c as

$$\mathcal{T}_{c,\gamma} = \{a_{c,1}, \ldots, a_{c,T_c}\}. \tag{132}$$

For the sample $\{(S_i, Y_i)\}_{i=1}^n$, the atypical and typical cell index sets are defined as

$$\mathcal{U}_c^0 = \{i \in [n] : Y_i = y_c,\ S_i \notin \mathcal{T}_{c,\gamma}\}, \tag{133}$$

$$\mathcal{U}_{c,k} = \{i \in [n] : Y_i = y_c,\ S_i = a_{c,k}\}, \qquad k \in [T_c]. \tag{134}$$

We define

$$p_c^0 = \mathbb{P}(Y = y_c, S \notin \mathcal{T}_{c,\gamma} \mid T=1), \quad p_{c,k} = \mathbb{P}(Y = y_c, S = a_{c,k} \mid T=1), \quad b_{c,k} = \widehat{\ell}_S(a_{c,k}, y_c). \tag{135}$$

The empirical surrogate risk admits the exact partition

$$\frac{1}{n}\sum_{i=1}^n \widehat{\ell}_S(S_i, Y_i) = \frac{1}{n}\sum_{c=1}^K \sum_{i \in \mathcal{U}_c^0} \widehat{\ell}_S(S_i, y_c) + \sum_{c=1}^K \sum_{k=1}^{T_c} \frac{|\mathcal{U}_{c,k}|}{n}\, b_{c,k}. \tag{136}$$

The population surrogate risk has the corresponding decomposition

$$\mathbb{E}_{(S,Y)\sim p(S,Y\mid T=1)}[\widehat{\ell}_S(S,Y) \mid T=1] = \sum_{c=1}^K p_c^0\, \mu_c^0 + \sum_{c=1}^K \sum_{k=1}^{T_c} p_{c,k} b_{c,k}, \tag{137}$$

The conditional surrogate loss on the atypical region is denoted by μ_c^0:

$$\mu_c^0 = \mathbb{E}_{(S,Y)\sim p(S,Y\mid T=1)}[\widehat{\ell}_S(S, y_c) \mid Y = y_c, S \notin \mathcal{T}_{c,\gamma}, T=1]. \tag{138}$$

Subtracting Eq. (136) from Eq. (137) separates the surrogate population–empirical discrepancy into an atypical-region contribution and a typical-cell contribution. For each class y_c, we denote the empirical average surrogate loss within the atypical region by $\widehat{\mu}_c^0$. When $|\mathcal{U}_c^0| > 0$, this quantity is defined by

$$\widehat{\mu}_c^0 = \frac{1}{|\mathcal{U}_c^0|} \sum_{i \in \mathcal{U}_c^0} \widehat{\ell}_S(S_i, y_c). \tag{139}$$

When $|\mathcal{U}_c^0| = 0$, we use the convention $\widehat{\mu}_c^0 = 0$. The difference between Eq. (137) and Eq. (136) can be written explicitly as the following three-term decomposition:

$$\begin{aligned}
&\mathbb{E}_{(S,Y)\sim p(S,Y\mid T=1)}[\widehat{\ell}_S(S,Y) \mid T=1] - \frac{1}{n}\sum_{i=1}^n \widehat{\ell}_S(S_i, Y_i) \\
&= \underbrace{\sum_{c=1}^K p_c^0(\mu_c^0 - \widehat{\mu}_c^0)}_{P_1} + \underbrace{\sum_{c=1}^K \left(p_c^0 - \frac{|\mathcal{U}_c^0|}{n}\right)\widehat{\mu}_c^0}_{P_2} + \underbrace{\sum_{c=1}^K \sum_{k=1}^{T_c} \left(p_{c,k} - \frac{|\mathcal{U}_{c,k}|}{n}\right) b_{c,k}}_{P_3}.
\end{aligned} \tag{140}$$

The quantity P_1 measures the conditional loss mismatch inside the atypical region under its population mass. The quantity P_2 measures the class-wise sampling error of the atypical mass. The quantity P_3 measures the empirical-frequency discrepancy over the typical representation cells.

For notational compactness, we define $M_S = \sup_{s,y} \widehat{\ell}_S(s,y)$ on the common activation envelope. Since $0 \le \widehat{\mu}_c^0 \le M_S$ and $0 \le \mu_c^0 \le M_S$, the first contribution term satisfies

$$\begin{aligned}
P_1 &= \sum_{c=1}^K p_c^0(\mu_c^0 - \widehat{\mu}_c^0) \le \sum_{c=1}^K p_c^0 \mu_c^0 \\
&\le M_S \sum_{c=1}^K p_c^0 = M_S \mathbb{P}(S \notin \mathcal{T}_{Y,\gamma} \mid T=1) \\
&\le M_S \frac{\gamma}{\sqrt{n}}.
\end{aligned} \tag{141}$$

The final inequality follows by averaging Eq. (126) over Y. This bound controls the part of the atypical-region discrepancy that can arise from the conditional surrogate loss itself.

The second contribution term has the exact form

$$P_2 = \sum_{c=1}^{K} \left(p_c^0 - \frac{|\mathcal{U}_c^0|}{n} \right) \widehat{\mu}_c^0. \tag{142}$$

The bound $0 \le \widehat{\mu}_c^0 \le M_S$ gives

$$P_2 \le M_S \sum_{c=1}^{K} \left| p_c^0 - \frac{|\mathcal{U}_c^0|}{n} \right|. \tag{143}$$

The remaining task is to control the class-wise atypical-frequency deviations. For each class y_c and sample index i, we define the Bernoulli variable

$$I_{i,c} = \mathbf{1}\{Y_i = y_c,\ S_i \notin \mathcal{T}_{c,\gamma}\}.$$

Its expectation and empirical average satisfy

$$\mathbb{E}[I_{i,c}] = p_c^0, \qquad \frac{1}{n}\sum_{i=1}^{n} I_{i,c} = \frac{|\mathcal{U}_c^0|}{n}.$$

The centered variable $Z_{i,c} = I_{i,c} - p_c^0$ satisfies $\mathbb{E}[Z_{i,c}] = 0$ and $|Z_{i,c}| \le 1$. Its variance satisfies

$$\mathrm{Var}(Z_{i,c}) = \mathrm{Var}(I_{i,c}) = p_c^0(1 - p_c^0) \le p_c^0.$$

Consequently, the total variance of the n centered variables is bounded by

$$\sum_{i=1}^{n} \mathrm{Var}(Z_{i,c}) \le n p_c^0.$$

For the independent centered variables $Z_{i,c}$ defined above, which satisfy $\mathbb{E}[Z_{i,c}] = 0$ and the uniform absolute bound $|Z_{i,c}| \le 1$, Bernstein's inequality yields, for every $x > 0$, the following two-sided concentration bound:

$$\mathbb{P}\left(\left| \frac{1}{n}\sum_{i=1}^{n} Z_{i,c} \right| > \sqrt{\frac{2p_c^0 x}{n}} + \frac{2x}{3n} \right) \le 2e^{-x}.$$

We set $x = \log(2K/\delta_1)$. For each fixed class y_c, the resulting failure probability is at most δ_1/K. A union bound over all K classes shows that, with probability at least $1 - \delta_1$, the following inequalities hold simultaneously for every $c \in [K]$:

$$\left| p_c^0 - \frac{|\mathcal{U}_c^0|}{n} \right| \le \sqrt{\frac{2p_c^0 \log(2K/\delta_1)}{n}} + \frac{2\log(2K/\delta_1)}{3n}. \tag{144}$$

Substituting these class-wise deviations into the preceding bound for P_2 gives

$$P_2 \le M_S \sqrt{\frac{2\log(2K/\delta_1)}{n}} \sum_{c=1}^{K} \sqrt{p_c^0} + \frac{2KM_S \log(2K/\delta_1)}{3n}. \tag{145}$$

The Cauchy–Schwarz inequality gives

$$\sum_{c=1}^{K} \sqrt{p_c^0} \le \sqrt{K \sum_{c=1}^{K} p_c^0}. \tag{146}$$

The class-uniform estimate in Eq. (126) implies

$$\sum_{c=1}^{K} p_c^0 = \mathbb{P}(S \notin \mathcal{T}_{Y,\gamma} \mid T = 1) \le \frac{\gamma}{\sqrt{n}}.$$

Combining the preceding two inequalities yields the following bound for the second contribution term with probability at least $1 - \delta_1$:

$$P_2 \le M_S \frac{\sqrt{2K\gamma \log(2K/\delta_1)}}{n^{3/4}} + \frac{2KM_S \log(2K/\delta_1)}{3n}. \tag{147}$$

The factor $n^{-3/4}$ arises from the composition of the atypical-mass scale $\sum_c p_c^0 = O(n^{-1/2})$ and the variance-sensitive Bernstein fluctuation of order $O(\sqrt{p_c^0/n})$.

Lemma 1 *Let $(X_1, \ldots, X_J)$ follow a multinomial distribution with m trials and cell probabilities $(p_1, \ldots, p_J)$. Let $\bar{a}_1, \ldots, \bar{a}_J \geq 0$ be fixed weights, and assume that $\sum_{j=1}^{J} \bar{a}_j p_j \neq 0$. We define the variance factor by*

$$\beta = 2 \sum_{j=1}^{J} \bar{a}_j^2 p_j.$$

For every $\eta > 0$, the following concentration inequality holds:

$$\mathbb{P}\left(\sum_{j=1}^{J} \bar{a}_j \left(p_j - \frac{X_j}{m} \right) > \eta \right) \leq \exp\left(-\frac{m\eta^2}{\beta} \right).$$

The remaining task is to control the typical-cell discrepancy. We denote this weighted frequency error by P_3, which is defined by the following expression:

$$P_3 = \sum_{c=1}^{K} \sum_{k=1}^{T_c} \left(p_{c,k} - \frac{|\mathcal{U}_{c,k}|}{n} \right) b_{c,k}. \tag{148}$$

Next, we define

$$d_{c,k} = p_{c,k} - \frac{|\mathcal{U}_{c,k}|}{n}, \qquad P_{3,c} = \sum_{k=1}^{T_c} d_{c,k} b_{c,k}. \tag{149}$$

This definition gives $P_3 = \sum_c P_{3,c}$. For each reference cell $k \in [T_c]$, we introduce the leave-one-cell discrepancy as

$$D_{c,k} = \sum_{t=1}^{T_c} d_{c,t} b_{c,t} - d_{c,k} b_{c,k} = \sum_{t \neq k} d_{c,t} b_{c,t}. \tag{150}$$

For each class y_c, the listed typical cells do not form an exhaustive partition of the sample space. Thus, we introduce a complementary category that collects all outcomes outside these cells. Its probability and empirical count are defined as

$$p_0 = 1 - \sum_{j=1}^{J} p_j, \qquad N_0 = n - \sum_{j=1}^{J} N_j.$$

The augmented count vector $(N_0, N_1, \ldots, N_J)$ is consequently multinomial with n trials and cell probabilities $(p_0, p_1, \ldots, p_J)$. The complementary category serves only to complete the multinomial partition. Accordingly, we assign it zero weight, i.e., $a_0 = 0$, which leaves both the weighted deviation and its variance factor unchanged:

$$a_0 \left(p_0 - \frac{N_0}{n} \right) + \sum_{j=1}^{J} a_j \left(p_j - \frac{N_j}{n} \right) = \sum_{j=1}^{J} a_j \left(p_j - \frac{N_j}{n} \right),$$

and

$$2a_0^2 p_0 + 2 \sum_{j=1}^{J} a_j^2 p_j = 2 \sum_{j=1}^{J} a_j^2 p_j.$$

Therefore, applying Lemma 1 with $m = n$, $X_j = N_j$, and $\bar{a}_j = a_j$ gives

$$\mathbb{P}\left(\sum_{j=1}^{J} a_j \left(p_j - \frac{N_j}{n} \right) > \eta \right) \leq \exp\left[-\frac{n\eta^2}{2 \sum_{j=1}^{J} a_j^2 p_j} \right]. \tag{151}$$

To control the leave-one-cell discrepancy $D_{c,k}$, we specialize Eq. (151) by choosing the deterministic weights $a_t = b_{c,t}$ for $t \neq k$ and setting $a_k = 0$, thereby excluding the reference cell k and matching the weighted deviation exactly to Eq. (150). Under this specialization, the variance-sensitive term governing the concentration rate in Eq. (151) becomes $2 \sum_{t \neq k} b_{c,t}^2 p_{c,t}$. Under the uniform constraint $0 \leq b_{c,t} \leq M_S$, this quantity admits the following upper estimate:

$$\begin{aligned} 2 \sum_{t \neq k} b_{c,t}^2 p_{c,t} &\leq 2M_S^2 \sum_{t \neq k} p_{c,t} \\ &\leq 2M_S^2 (p_c - p_{c,k}), \end{aligned} \tag{152}$$

Here, $p_c = \mathbb{P}(Y = y_c \mid T = 1)$. The resulting concentration bound holds for every reference cell k:

$$\mathbb{P}\left(D_{c,k} > M_S \sqrt{\frac{2(p_c - p_{c,k}) \log(1/\delta')}{n}} \right) \leq \delta'. \tag{153}$$

We choose $\delta' = \delta_2/(KT_c)$ and apply a union bound over all classes and all typical cells. With probability at least $1 - \delta_2$, Eq. (153) holds simultaneously for every admissible pair (c, k). The union bound yields the following simultaneous inequality:

$$D_{c,k} \leq M_S \sqrt{\frac{2(p_c - p_{c,k}) \log(KT_c/\delta_2)}{n}}. \tag{154}$$

The leave-one-cell construction allows the original class discrepancy to be recovered exactly. Summing Eq. (150) over all reference cells yields

$$\begin{aligned} \sum_{k=1}^{T_c} D_{c,k} &= T_c P_{3,c} - \sum_{k=1}^{T_c} d_{c,k} b_{c,k} \\ &= (T_c - 1) P_{3,c}. \end{aligned} \tag{155}$$

For $T_c \geq 2$, Eqs. (154) and (155) imply

$$P_{3,c} \leq \frac{M_S}{T_c - 1} \sqrt{\frac{2 \log(KT_c/\delta_2)}{n}} \sum_{k=1}^{T_c} \sqrt{p_c - p_{c,k}}. \tag{156}$$

The concavity of the square-root function gives

$$\frac{1}{T_c} \sum_{k=1}^{T_c} \sqrt{p_c - p_{c,k}} \leq \sqrt{p_c - \frac{1}{T_c} \sum_{k=1}^{T_c} p_{c,k}} \leq \sqrt{p_c}. \tag{157}$$

Since $T_c/(T_c - 1) \leq 2$ for $T_c \geq 2$, the class-wise discrepancy obeys

$$P_{3,c} \leq 2 M_S \sqrt{p_c} \sqrt{\frac{2 \log(KT_c/\delta_2)}{n}}. \tag{158}$$

When $T_c = 1$, the same scaling follows directly from the two-category multinomial concentration between the single typical cell and its complement. Summing Eq. (158) over the classes gives

$$P_3 \leq 2\sqrt{2}\, M_S \sum_{c=1}^{K} \sqrt{p_c} \sqrt{\frac{\log(KT_c/\delta_2)}{n}}. \tag{159}$$

Applying the Cauchy–Schwarz inequality to the class-wise sum in Eq. (159) yields

$$P_3 \leq 2\sqrt{2K}\, M_S \sqrt{\frac{\sum_{c=1}^{K} p_c \log T_c + \log(K/\delta_2)}{n}}. \tag{160}$$

The typical-set cardinality in Eq. (131) implies

$$\begin{aligned} \sum_{c=1}^{K} p_c \log T_c &\leq \sum_{c=1}^{K} p_c H(S \mid Y = y_c, T = 1) + \epsilon_S \\ &= H(S \mid Y, T = 1) + \epsilon_S. \end{aligned} \tag{161}$$

Substitution of Eq. (161) into Eq. (160) yields

$$P_3 \leq 2\sqrt{2K}\, M_S \sqrt{\frac{H(S \mid Y, T = 1) + \epsilon_S + \log(K/\delta_2)}{n}}. \tag{162}$$

The combination of Eqs. (122), (141), (147), and (162) with $\delta_1 = \delta_2 = \delta/2$ gives

$$\begin{aligned} \mathrm{Gen}_n(S) \leq \rho_S + A_1 \frac{\gamma}{\sqrt{n}} + A_2 \frac{\sqrt{\gamma \log(4K/\delta)}}{n^{3/4}} + A_4 \frac{\log(4K/\delta)}{n} \\ + A_3 \sqrt{\frac{H(S \mid Y, T = 1) + \Gamma_S}{n}}, \end{aligned} \tag{163}$$

This choice gives $\log(2K/\delta_1) = \log(4K/\delta)$ and $\log(K/\delta_2) = \log(2K/\delta)$. The constants in Eq. (163) are $A_1 = M_S$, $A_2 = M_S\sqrt{2K}$, $A_3 = 2\sqrt{2K}\,M_S$, and $A_4 = 2KM_S/3$. The auxiliary quantity is $\Gamma_S = \epsilon_S + \log(2K/\delta)$, where ϵ_S is the concentration radius defined in Eq. (111). This bound establishes the first assertion of the theorem.

The information-theoretic form of the bound follows from the chain rule

$$I(S;Y \mid T=1) = H(S \mid T=1) - H(S \mid Y,T=1). \tag{164}$$

A rearrangement of Eq. (164) gives

$$H(S \mid Y,T=1) = H(S \mid T=1) - I(S;Y \mid T=1). \tag{165}$$

The representation-dependent part of Eq. (163) takes the following form

$$\mathcal{B}_n(S) = A_3\sqrt{\frac{H(S \mid T=1) - I(S;Y \mid T=1) + \Gamma_S}{n}}. \tag{166}$$

Eq. (166) makes the statistical role of visual retracing explicit. The population guarantee improves when the intervention increases target information more rapidly than it increases the total variability and sensitivity of the representation.

We now compare the baseline and retraced representations. Eqs (115)–(117) give

$$\begin{aligned} V_0 - V_1 &= \Big(I(S_1;Y \mid T=1) - I(S_0;Y \mid T=1)\Big) \\ &\quad - \Big(H(S_1 \mid T=1) - H(S_0 \mid T=1)\Big) - (\Gamma_{S_1} - \Gamma_{S_0}) \\ &= G_{\mathrm{VR}} - \Delta H - \Delta\Gamma. \end{aligned} \tag{167}$$

Whenever the right-hand side of Eq. (167) is positive, the difference of the two square-root complexity terms can be rationalized exactly:

$$\begin{aligned} \mathcal{B}_n(S_0) - \mathcal{B}_n(S_1) &= \frac{A_3}{\sqrt{n}}(\sqrt{V_0} - \sqrt{V_1}) \\ &= \frac{A_3}{\sqrt{n}}\frac{V_0 - V_1}{\sqrt{V_0} + \sqrt{V_1}} \\ &= \frac{A_3}{\sqrt{n}}\frac{G_{\mathrm{VR}} - \Delta H - \Delta\Gamma}{\sqrt{V_0} + \sqrt{V_1}}. \end{aligned}$$

The remaining lower-order terms use the same common boundedness envelope for both representations. The only representation-specific additive term that remains outside the square root is the surrogate residual. Hence, the complete upper bound is strictly smaller after retracing whenever

$$\rho_{S_1} - \rho_{S_0} < \frac{A_3}{\sqrt{n}}\frac{G_{\mathrm{VR}} - \Delta H - \Delta\Gamma}{\sqrt{V_0} + \sqrt{V_1}}. \tag{168}$$

Eq. (168) proves the final assertion of the theorem.

The quantity G_{VR} in Eq. (105) directly measures the change in target-relevant information between the baseline representation S_0 and the retraced representation S_1. A positive value therefore indicates that the retraced representation carries more information about the faithful target token than the baseline representation.

The scale-matching rule in Eq. (103) supplies an additional structural control that is absent from an unconstrained reinjection. The aligned retrieval satisfies

$$\|R\|_1 = \frac{\|\widetilde{f}\|_1}{\|R_0\|_1 + d\epsilon}\|R_0\|_1 \le \|\widetilde{f}\|_1, \tag{169}$$

The convex fusion satisfies the following norm bound.

$$\|\widehat{f}\|_1 \le (1-\lambda)\|\widetilde{f}\|_1 + \lambda\|R\|_1 \le \|\widetilde{f}\|_1, \tag{170}$$

$$\|\widehat{f} - \widetilde{f}\|_1 = \lambda\|R - \widetilde{f}\|_1 \le 2\lambda\|\widetilde{f}\|_1. \tag{171}$$

Eqs. (169)–(171) guarantee that the retracing branch does not enlarge the local norm envelope as a consequence of numerical scale mismatch in the visual memory. The scale-matching rule therefore controls the variability of the retraced representation and keeps the Lipschitz surrogate constant M_S under the same activation envelope. This control limits the representation-dependent entropy and sensitivity costs as well as the surrogate-residual increase that compete with G_{VR} in Eq. (168). This completes the derivation. □

C.3 INTERPRETATION FOR THE PAPER

The theorem yields a unified statistical explanation for the benefit of visual retracing. The representation encountered at a high-uncertainty layer can suffer from diminished visual grounding since the original visual tokens have been progressively processed without fresh access to the image. The retracing branch reopens that access and produces a query-conditioned vector whose coefficients $\langle h, z_j \rangle$ depend on the current decoding state. The quantity G_{VR} in Eq. (105) measures the resulting change in target-relevant information from S_0 to S_1. Therefore, the theorem requires this recovered information to be sufficiently large relative to the accompanying representation-complexity costs rather than assuming an unconditional benefit from visual reinjection.

The generalization part of the theorem then explains why this information gain has a population-level consequence. The class-conditional typical set contains at most $\exp(H(S \mid Y, T = 1) + \epsilon_S)$ high-probability representation states. A representation that is more predictive of Y decreases $H(S \mid Y, T = 1)$ through the identity $H(S \mid Y, T = 1) = H(S \mid T = 1) - I(S; Y \mid T = 1)$. Therefore, the empirical sample needs to estimate fewer effective high-probability representation cells when the increase in target information dominates the increase in representation variability. The multinomial concentration argument converts that reduction in effective cell count into a tighter population-empirical bound. In this sense, the theorem connects the image-specific information recovered by visual retracing to the statistical reliability of token prediction rather than merely to a local change in hidden-state magnitude.

The resulting sufficient condition is $G_{\mathrm{VR}} > \Delta H + \Delta\Gamma$. The complete guarantee additionally requires that the induced decrease in the square-root complexity term dominate any increase in the surrogate residual. Query-conditioned retrieval is favorable since it emphasizes visual tokens that are simultaneously aligned with the current hidden query and with the target decision. Scale matching is favorable since it prevents the recovered evidence from expanding the activation envelope. Uncertainty-triggered activation is favorable since it restricts intervention to contexts in which the baseline has weak visual reliance. One-shot retracing is favorable since it avoids accumulating multiple independent complexity costs after the first visual recovery event. Thus, the theorem provides a unified justification for the particular design of the visual retracing module instead of only proving that generic visual reinjection may help.

D INTERPRETATION AND ELUCIDATION OF TWO THEORETICAL PERSPECTIVES

Remark 1 *Theorem 1 shows that channel-wise evidence rebalancing improves faithful prediction by jointly enhancing supportive channels and suppressing adverse ones. The resulting increase in $A_\star$ directly enlarges the guaranteed margin gain $\xi(u)$ in Eq.* (14)*, while uncertainty-aware scaling further strengthens this effect in decoding states. Consequently, the correction reduces the mass of low-margin contexts, and whenever this reduction dominates the correction-failure and complexity terms in Eq.* (17)*, the local margin gain translates into a strictly tighter population-level token-selection guarantee. Theorem 2 shows that uncertainty-triggered visual retracing improves visually grounded prediction by increasing the target-relevant information in the internal representation. When $G_{\mathrm{VR}} > \Delta H + \Delta\Gamma$, the representation-dependent complexity term is strictly reduced. If this reduction further dominates the increase in the surrogate residual, the overall population-level generalization bound becomes tighter. The scale-matching mechanism further controls the retracing perturbation through $\|R\|_1 \le \|\widetilde{f}\|_1$, $\|\widehat{f}\|_1 \le \|\widetilde{f}\|_1$, $\|\widehat{f} - \widetilde{f}\|_1 \le 2\lambda\|\widetilde{f}\|_1$, thereby controlling the additional complexity introduced by visual retracing. Together, the increased target-relevant information and controlled representation complexity provide a theoretical explanation for improved visual grounding and reduced hallucination.*

E ALGORITHM WORKFLOW

The overall workflow of the proposed VERGE framework is summarized in Algorithm 1.

Algorithm 1: The proposed VERGE for Hallucination Mitigation in LVLMs

Require: Image v, textual prefix x, pre-trained MLLM $\mathcal{F}_\theta$, partition ratio ρ, correction hyperparameters α, β, γ.
Output: Generated response $y = (y_1, \ldots, y_T)$.

Encode the image v into visual tokens $\{z_1, \ldots, z_{N_v}\}$ and take the textual prefix x as input to $\mathcal{F}_\theta$.
Compute the magnitude of each visual token $r_i = \|z_i\|_2$ $(i = 1, \ldots, N_v)$, and determine the partition threshold $\theta = r_{(M)}$.
Split visual tokens into the high-magnitude subset $H = \{i : r_i \geq \theta\}$ and the evidence subset $E = \{1, \ldots, N_v\} \setminus \mathcal{H}$.
Initial set to trigger = TRUE;
for $t = 1, \ldots, T$ **do**
  for $l = l_e, \ldots, l_s$ **do**
    Obtain the hidden state h of the last token at each layer l.
    Compute the cosine similarity between h and each visual token z_i by Eq. (3), and obtain the normalized relevance weights π_i^H and π_i^E.
    Compute the pool-specific visual prototypes c^H and c^E.
    Compute the channel-wise similarities b_j^H and b_j^E by Eq. (4).
    Compute the channel-wise suppression mask m and enhancement bias g by Eq. (5). ;
    Modulate the fine-grained correction strength by Eq. (8).
    Apply fine-grained FFN correction to the intermediate state a by Eq. (6).
  end
  Apply the subsequent layer-wise and adaptive decoding corrections based on the updated representations.
  Generate the next token y_t using $\mathcal{F}_\theta$ with the corrected representations.
  Update the decoding context with y_t.
end
for $l = 1, \ldots, L$ **do**
  Compute predictive uncertainty u by Eq. (7);
  if $\text{trigger} = \text{TRUE}$ *and* $u > u_0$ **then**
    Compute two scaling coefficients s_1 and s_2 by Eq. (9);
    Execute visual evidence rejection by Eq. (10).
  end
end
Obtain the image-conditioned distribution ℓ_t^+ and the text-only distribution ℓ_t^- by Eq. (17).
Compute the text-prior contrastive score $\tilde{\ell}_t$ by Eq. (18).
Compute the adaptive plausibility ratio $\kappa(\delta)$ by Eq. (19).
Construct the plausible candidate set $\mathcal{V}$ by Eq. (20).
Compute the final token distribution p_{uni} by Eq. (21) and sample y_t.
return $y = (y_1, \ldots, y_T)$.

F DETAILED DESCRIPTION OF BENCHMARKS AND METRICS

CHAIR evaluates object hallucination in free-form image descriptions. It extracts object mentions from a generated caption and checks them against the image-level annotations. In a commonly used configuration, evaluation draws on 500 images sampled from MS-COCO (Lin et al., 2014), which provides annotations for 80 object categories. The sentence-level hallucination rate is denoted by C_S and the instance-level rate by C_I:

$$C_S = \frac{\#\{\text{captions containing an unsupported object}\}}{\#\{\text{captions}\}}, \qquad C_I = \frac{\#\{\text{unsupported object mentions}\}}{\#\{\text{all object mentions}\}}.$$

Lower values of C_S and C_I indicate fewer hallucinated objects. We may also report category-level object precision, namely the fraction of distinct generated object categories that are supported by the annotations. Strictly speaking, CHAIR is a caption-level metric and evaluation protocol built on COCO images, rather than a standalone image collection.

POPE recasts object hallucination as binary visual question answering (Li et al., 2023). Given an image and a queried object, the model must answer whether that object is present. Negative examples are produced with three sampling schemes: random sampling selects absent objects without an additional constraint, popular sampling emphasizes frequently occurring object categories, and adversarial sampling chooses absent objects that are likely to co-occur with objects visible in the image. We report Accuracy and F1 for each split and their average, so that performance is compared under several levels of negative-sample difficulty.

HallusionBench is designed to expose failures caused by visual illusions and conflicts between image evidence and language priors (Guan et al., 2024). Its examples pair images with questions that are intentionally ambiguous, deceptive, or knowledge-conflicting, requiring the model to ground its answer in the actual visual content. We use figure-level accuracy (fACC), question-level accuracy (qACC), easy- and hard-subset accuracy (easyA and hardA), and answer-level accuracy (aACC). These complementary scores summarize performance at the figure, question, and answer levels.

LLaVA-Bench evaluates open-ended visual instruction following on challenging natural images (Liu et al., 2023). Its 60 questions span conversational interaction, detailed image description, and complex visual reasoning. Because the responses are free-form, a reference-aware language-model judge is used to compare the generated answers with the provided reference answers; in the source evaluation, this role is assigned to GPT-5.2. The benchmark therefore checks whether improved visual grounding is accompanied by preserved descriptive completeness, instruction following, and reasoning quality.

MME adopts a unified yes-or-no evaluation protocol and separates multimodal ability into perception and cognition (Fu et al., 2026). Its perception suite covers existence, counting, position, color, posters, celebrities, scenes, landmarks, artwork, and OCR. The cognition suite includes commonsense reasoning, numerical calculation, text translation, and code reasoning. The resulting 14-task collection allows object- and attribute-level perception to be examined alongside higher-level reasoning, making it useful for checking whether hallucination reduction damages broader model competence.

MM-Vet measures integrated capability through open-ended visual questions (Yu et al., 2023). Individual questions can require several skills at once, including visual recognition, OCR, external knowledge, spatial reasoning, language generation, and mathematical reasoning. The benchmark consequently probes how well these abilities are coordinated in one response, rather than scoring each skill in isolation. It is particularly suitable for testing whether an inference-time intervention keeps multimodal reasoning effective while reducing unsupported visual claims.

VizWiz contains visual questions collected from blind users in everyday settings Gurari et al. (2018). The accompanying photographs often suffer from blur, occlusion, poor framing, weak illumination, ambiguous content, or insufficient evidence for a confident answer. These properties make the benchmark a demanding test of visual robustness and uncertainty handling: a capable model should use the available evidence while avoiding confident guesses when the image does not support a conclusion.

MMBench is a multiple-choice benchmark organized by a fine-grained taxonomy of visual perception and reasoning abilities (Liu et al., 2024d). It applies circular evaluation, which permutes the answer options across equivalent versions of a question to reduce positional bias and expose unstable predictions. The objective scoring procedure provides a standardized complement to open-ended benchmarks and enables a controlled assessment of whether the model retains broad visual-language competence after hallucination mitigation.

G Additional Empirical Findings

Mechanism Analysis of Contrastive Decoding. Fig. 7 provides a detailed analysis of how the proposed language-prior-calibrated contrastive decoding reshapes the candidate distribution. We

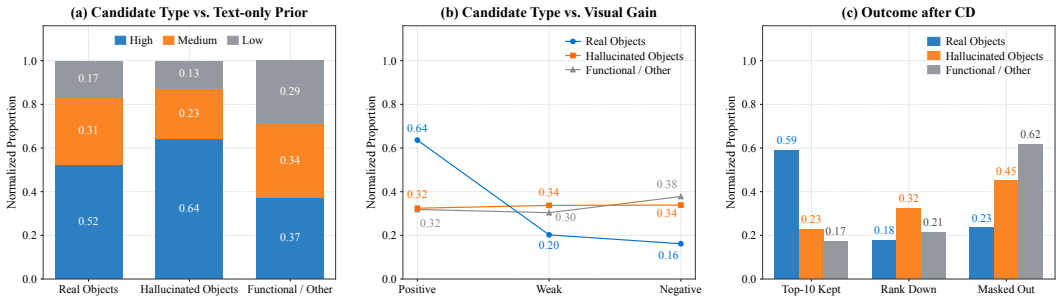


Figure 7: Left: text-only prior distributions of different candidate types. Middle: visual gains induced by image conditioning. Right: candidate outcomes after contrastive re-ranking and plausibility filtering.

first fix the decoding trajectory without contrastive decoding and collect the original top-10 candidates from the image-conditioned branch at each word-start decoding position. The same candidate records are then tracked throughout the three panels. According to the ground-truth object annotations, object candidates that are present in the image are categorized as *Real Objects*, object candidates that are absent from the image are categorized as *Hallucinated Objects*, and the remaining candidates, including functional tokens and other unmatched tokens, are grouped as *Functional / Other*. Here, "Hallucinated Objects" refers to visually unsupported object candidates in the candidate pool rather than hallucinations that have already been generated.

The left panel characterizes the language-prior preference of these candidates using the text-only branch. Specifically, all vocabulary candidates are ranked according to their text-only log-probability $\ell^-(w)$. Candidates ranked within the top 10 are assigned to the *High* group, those ranked from 11 to 50 are assigned to the *Medium* group, and the remaining candidates are assigned to the *Low* group. The proportions are normalized independently within each candidate type. As shown in the left panel, 64% of hallucinated-object candidates belong to the high-prior group, compared with 52% of real-object candidates. Meanwhile, only 13% of hallucinated-object candidates fall into the low-prior group. This distribution indicates that visually unsupported object candidates are more likely to receive strong preference from the textual context alone, revealing a clear source of language-prior-induced hallucination.

The middle panel further examines how visual conditioning changes the preference for the same candidates. Since $\ell^+(w)$ and $\ell^-(w)$ denote the image-conditioned and text-only log-probabilities, respectively, the visual support for a candidate can be directly characterized by

$$\ell^+(w) - \ell^-(w).$$

Accordingly, we divide the candidates into three groups:

$$\begin{cases} \ell^+(w) - \ell^-(w) > 0.1, & \text{Positive}, \\ -0.1 \le \ell^+(w) - \ell^-(w) \le 0.1, & \text{Weak}, \\ \ell^+(w) - \ell^-(w) < -0.1, & \text{Negative}. \end{cases}$$

The middle panel shows a clear distinction between grounded and unsupported object candidates. In particular, 64% of real-object candidates receive positive visual gain, whereas this proportion decreases to only 32% for hallucinated-object candidates. Conversely, 34% of hallucinated-object candidates exhibit negative visual gain, compared with only 16% of real-object candidates. These results suggest that image conditioning provides substantially stronger evidence for grounded objects, while many hallucination-prone candidates are primarily supported by the textual context rather than the visual input. This discrepancy is explicitly exploited by our contrastive decoding score

$$\tilde{\ell}(w) = (1+\delta)\ell^+(w) - \delta\ell^-(w) = \ell^+(w) + \delta\big(\ell^+(w) - \ell^-(w)\big).$$

The second form makes the mechanism particularly clear: the original image-conditioned score $\ell^+(w)$ is adjusted according to the additional evidence contributed by the image. Therefore, candidates whose probabilities are increased by visual conditioning receive a positive correction,

whereas candidates whose preference mainly originates from the text-only prior are relatively suppressed. However, contrastive reweighting alone may undesirably promote low-probability candidates with a large relative visual gain. Therefore, we further apply the adaptive plausibility constraint $\kappa(\delta) = \kappa_0 \cdot \min(1, (\delta/\delta_{\text{ref}})^q)$, and construct the plausible candidate set as $\mathcal{V} = \{w \mid p_\theta(w \mid v, x, y_{<t}) \geq \kappa(\delta) \max_{w'} p_\theta(w' \mid v, x, y_{<t})\}$. For the diagnostic analysis in Fig. 7, we use $\delta = 1$, $\kappa_0 = 0.1$, $\delta_{\text{ref}} = 1$, and $q = 2$, which gives $\kappa(\delta) = 0.1$.

The right panel finally traces the outcomes of the original Top-10 candidates after applying both contrastive re-ranking and the plausibility constraint. A candidate is categorized as *Top-10 Kept* if it survives the plausibility constraint, i.e., $w \in \mathcal{V}$, and remains among the top-10 candidates ranked by $\tilde{\ell}(w)$; it is categorized as *Rank Down* if it survives the constraint but falls outside the top 10 after re-ranking; otherwise, it is categorized as *Masked Out* when $w \notin \mathcal{V}$. As shown in the right panel, 59% of real-object candidates remain in the top 10 after contrastive decoding, whereas only 23% of hallucinated-object candidates are retained. In contrast, 32% of hallucinated candidates are pushed out of the top 10 and another 45% are directly removed by the plausibility constraint. In other words, 77% of the hallucination-prone candidates are either demoted or filtered after contrastive decoding, while visually grounded candidates are more likely to preserve their high-ranking positions.

Taken together, the three panels reveal a coherent mechanism behind the proposed contrastive decoding. Hallucination-prone object candidates tend to exhibit stronger text-only priors but receive substantially weaker support from image conditioning. The contrastive term exploits this discrepancy to reduce their relative ranking, while the plausibility constraint further removes candidates that remain insufficiently supported by the image-conditioned distribution. Consequently, the surviving candidate distribution is shifted toward visually grounded objects, thereby reducing language-prior dominance without indiscriminately suppressing visually supported predictions.

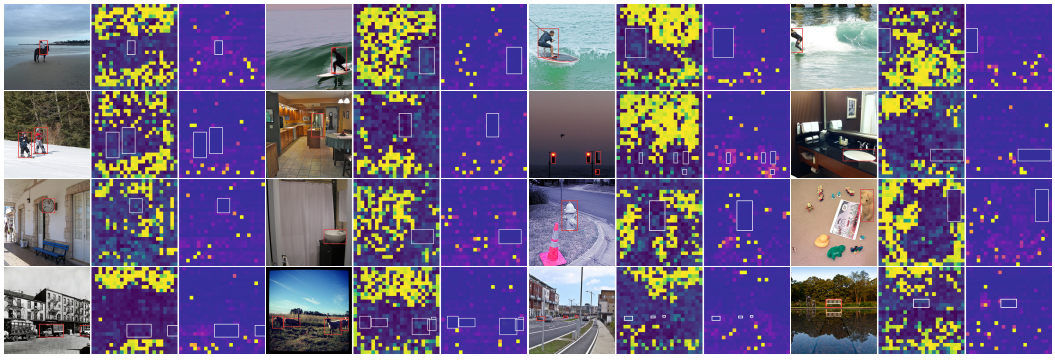


Figure 8: Each triptych shows the COCO image, visual-token norms, and visual attention from the first subword token of a grounded object mention to the visual tokens. Warmer colors indicate larger norms or stronger visual attention.

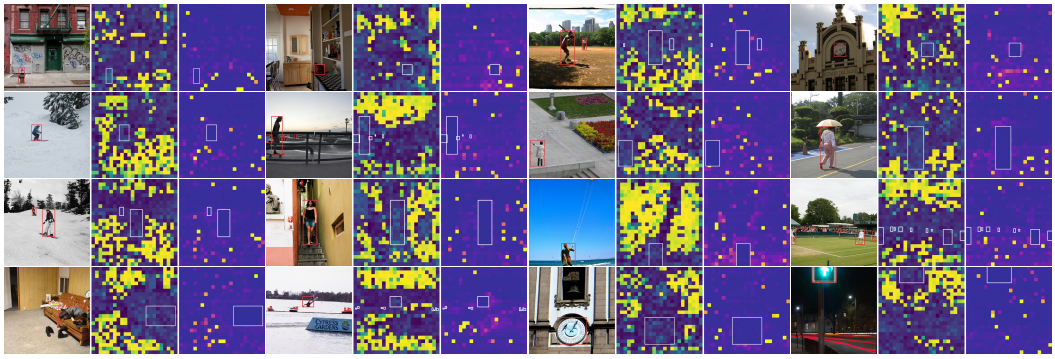


Figure 9: Each triptych shows the COCO image, visual-token norms, and visual attention from the first subword token of a grounded object mention to the visual tokens. Warmer colors indicate larger norms or stronger visual attention.

More qualitative examples are provided in Figs. 8 and 9, further illustrating the recurrent presence of high-magnitude visual tokens and their spatial misalignment with object-discriminative regions. These examples consistently support the observations in Fig. 1, demonstrating that unusually strong visual responses can divert attention away from target-relevant visual evidence across diverse images and object categories.

For completeness, we provide the detailed construction of Fig. 3. In the left panel, generated tokens are grouped into real-object, hallucinated-object, attribute/relation, and function tokens according to their normalized lexical categories and the image-level ground-truth object set. Following the notation in Sec. 3.1, visual tokens are partitioned into the high-magnitude subset R and the complementary evidence subset E according to $r_i = \|z_i\|_2$. This partition is determined from the projected visual representations and remains fixed across decoder layers. For generated token t at layer l, let $h_t^{(l)}$ and $z_i^{(l)}$ denote the corresponding textual and visual hidden states, with bars indicating ℓ_2 normalization as in Eq. (4). We compute the High-Norm Alignment Gap as

$$G_t^{(l)} = \frac{1}{|R|} \sum_{i \in R} \left\langle \bar{h}_t^{(l)}, \bar{z}_i^{(l)} \right\rangle - \frac{1}{|E|} \sum_{i \in E} \left\langle \bar{h}_t^{(l)}, \bar{z}_i^{(l)} \right\rangle .$$

Thus, a positive value indicates that the current textual representation is more strongly aligned with high-magnitude visual tokens, whereas a negative value indicates stronger alignment with the remaining visual evidence. The resulting scores are aggregated by token category over consecutive four-layer bins to obtain the left-panel statistics. In the middle panel, visual tokens are ranked within each image by r_i, and either the highest- or lowest-norm $p\%$ of tokens are selected for intervention, with $p \in \{5, 10, 15, 20, 25, 30\}$. The selected token representations are then replaced with the mean visual representation computed over all visual tokens from the same image, while all unselected visual tokens and the remaining inference configuration are left unchanged. The model subsequently regenerates the responses, and the resulting changes in C_I and object Recall relative to the original generation are reported to directly compare the functional consequences of perturbing high- and low-magnitude visual representations. In the right panel, ground-truth object boxes are obtained from the corresponding MS-COCO instance annotations and aligned with the visual patch layout after applying the same image preprocessing used by the model. Each patch is subsequently labeled according to whether it falls inside or outside the annotated target regions. Visual tokens are ranked by norm within each image and divided into ten percentile intervals from $0-10$ to $90-100$, after which we compute the composition of inside and outside patches within each interval. This analysis characterizes the spatial distribution of visual evidence across the token-magnitude spectrum and reveals the progressively reduced overlap between high-magnitude tokens and ground-truth object regions.

H Additional Experiential Result

H.1 Comparison Result

Table 3: Performance evaluation on CHAIR benchmark. The best results are highlighted in bold.

Methods	LLaVA-1.5-7B			LLaVA-1.5-13B			Qwen2.5-VL-7B			InternVL3-8B			GLM-4V-9B			Avg.	
	$C_S\downarrow$	$C_I\downarrow$	R↑	$C_S\downarrow$	$C_I\downarrow$	R↑	$C_S\downarrow$	$C_I\downarrow$	R↑	$C_S\downarrow$	$C_I\downarrow$	R↑	$C_S\downarrow$	$C_I\downarrow$	R↑	$C_S\downarrow$	$C_I\downarrow$
Decoding Strategy																	
Greedy	52.8	15.7	77.4	51.0	14.6	77.2	39.8	10.3	69.2	30.6	8.5	69.2	43.6	9.1	73.1	43.6	11.6
OPERA	49.0	13.4	70.3	50.8	15.1	78.4	36.0	12.2	58.7	32.0	9.0	69.4	39.6	9.3	68.0	41.5	11.8
Contrastive Decoding																	
VCD	51.4	15.2	76.9	50.0	13.2	76.7	41.2	10.7	68.5	33.6	9.8	70.0	43.9	9.3	72.0	44.0	11.6
ICD	54.6	16.1	72.8	54.4	15.6	74.7	41.8	10.6	68.8	30.6	8.3	68.0	44.3	9.4	72.4	45.1	12.0
Visual Intervention																	
MemVR	46.6	13.0	80.8	50.6	14.5	77.0	36.6	9.9	69.1	30.3	8.4	68.6	42.4	8.9	72.6	41.3	10.9
Ours	**45.8**	**12.8**	**81.1**	**42.6**	**12.9**	**79.2**	**32.1**	**8.9**	**69.4**	**30.1**	**8.1**	**70.2**	**38.9**	**8.5**	**73.4**	**37.9**	**10.2**
Δ%	↓1.7%	↓1.5%	↑0.4%	↓14.9%	↓2.6%	↑1.0%	↓10.8%	↓10.1%	↑0.3%	↓0.7%	↓2.4%	↑0.3%	↓1.8%	↓4.5%	↑0.4%	↓6.0%	↓4.2%

Table 4: POPE hallucination evaluation on five models. The best results are highlighted in bold.

Methods	Random		Popular		Adversarial		Average	
	Accuracy ↑	F1 ↑	Accuracy ↑	F1 ↑	Accuracy ↑	F1 ↑	Accuracy ↑	F1 ↑
Qwen2.5-VL-7B								
Greedy	88.50	87.16	87.63	86.32	86.83	85.57	87.66	86.35
ICD	89.17	88.02	88.00	86.90	86.97	85.95	88.04	86.96
VCD	89.53	88.51	88.53	87.55	87.03	86.13	88.37	87.40
MemVR	89.60	88.59	88.50	87.54	**87.57**	**86.67**	88.56	87.60
PAI	89.40	88.27	88.27	87.18	87.17	86.15	88.28	87.20
ML	86.13	83.99	85.43	83.31	84.93	82.83	85.50	83.38
VISTA	90.70	89.75	87.33	86.23	85.53	83.99	87.86	86.65
Ours	**92.00**	**91.30**	**88.57**	**87.96**	86.77	85.73	**89.11**	**88.33**
InternVL3-8B								
Greedy	**93.47**	93.10	90.67	90.47	88.17	88.21	90.77	90.59
ICD	93.33	93.04	90.70	90.37	88.50	88.48	90.84	90.63
VCD	93.43	**93.29**	90.17	89.96	88.27	88.43	90.62	90.56
MemVR	93.30	92.95	**91.07**	**90.85**	88.30	88.34	90.89	90.71
PAI	93.17	92.73	90.50	90.34	88.33	88.28	90.67	90.45
ML	93.10	92.65	90.60	90.27	87.87	87.82	90.52	90.25
VISTA	92.33	91.88	90.67	90.28	88.30	88.11	90.43	90.09
Ours	93.17	92.79	91.03	90.75	**88.70**	**88.61**	**90.97**	**90.72**
GLM-4V-9B								
Greedy	91.97	91.45	88.87	88.49	87.17	87.00	89.33	88.98
ICD	92.30	91.84	89.37	89.04	87.67	87.54	89.78	89.47
VCD	92.50	92.07	89.37	89.12	**87.97**	**87.91**	89.94	89.70
MemVR	92.67	92.25	89.40	89.17	87.73	87.68	89.93	89.70
PAI	92.40	91.93	89.07	88.79	87.40	87.30	89.62	89.34
ML	92.50	91.97	88.83	88.56	87.17	87.07	89.50	89.20
VISTA	92.93	92.55	89.47	89.29	87.60	87.62	90.00	89.82
Ours	**93.00**	**92.63**	**89.53**	**89.36**	87.80	87.82	**90.11**	**89.94**
LLaVA-1.5-7B								
Greedy	86.10	84.58	84.30	82.87	81.97	80.80	84.12	82.75
ICD	88.03	86.90	85.97	84.98	82.67	82.09	85.56	84.66
OPERA	86.30	84.60	85.10	83.40	82.90	81.30	84.77	83.10
VCD	88.43	87.40	86.37	85.48	83.07	82.59	85.96	85.16
MemVR	88.57	87.53	**87.10**	86.15	84.80	**84.04**	86.82	85.91
PAI	87.73	86.78	86.87	85.95	84.17	83.47	86.26	85.40
ML	78.17	80.24	81.77	82.94	74.47	77.63	78.13	80.27
VISTA	89.90	89.34	86.80	86.64	84.80	83.64	87.17	86.54
Ours	**90.13**	**89.45**	87.00	**86.75**	**85.03**	83.75	**87.39**	**86.65**
LLaVA-1.5-13B								
Greedy	**91.13**	90.27	88.00	86.62	81.90	81.97	87.01	86.28
ICD	89.63	88.76	87.57	86.65	84.27	84.06	87.16	86.49
OPERA	90.00	89.20	87.80	87.20	84.70	84.50	87.50	86.97
VCD	89.73	88.56	88.63	87.43	82.93	83.34	87.10	86.44
MemVR	89.43	89.20	88.90	88.40	**86.90**	**86.50**	88.41	88.03
PAI	88.87	87.47	86.43	84.83	83.27	82.02	86.19	84.77
ML	87.73	87.28	85.93	85.62	82.23	82.46	85.30	85.12
VISTA	90.33	90.21	89.00	89.11	86.50	85.71	88.61	88.34
Ours	90.37	**90.38**	**89.10**	**89.28**	86.57	85.88	**88.68**	**88.51**

Table 3 and 4 report the results on CHAIR and POPE across five MLLM backbones. On CHAIR, VERGE consistently achieves the lowest hallucination rates, reducing the average C_S and C_I to 37.9 and 10.2, with

relative improvements of 6.0% and 4.2% over the strongest competing results. The gains are particularly pronounced on Qwen2.5-VL-7B, where C_S and C_I decrease by 10.8% and 10.1%, respectively. Meanwhile, object recall is consistently preserved or improved, indicating that the reduction in hallucination is not achieved through conservative generation. On POPE, VERGE further demonstrates robust performance across different object-existence polling settings. Although a few competing methods achieve slightly higher scores on individual Random, Popular, or Adversarial splits, VERGE attains the best averaged Accuracy and F1 on all five backbones, indicating a more balanced improvement across different negative-sampling strategies. Specifically, VERGE achieves average Accuracy/F1 scores of 89.11/88.33, 90.97/90.72, 90.11/89.94, 87.39/86.65, and 88.68/88.51 on Qwen2.5-VL-7B, InternVL3-8B, GLM-4V-9B, LLaVA-1.5-7B, and LLaVA-1.5-13B, respectively. Averaged across the five backbones, VERGE reaches 89.25 Accuracy and 88.83 F1, outperforming the strongest competing method by 0.33 and 0.44 points, respectively. These results suggest that VERGE does not rely on gains from a particular evaluation setting, but provides consistently reliable hallucination mitigation across heterogeneous model architectures and different sampling strategies.

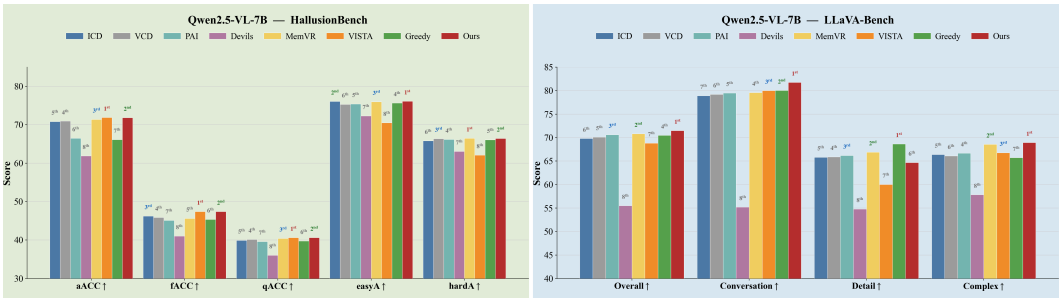
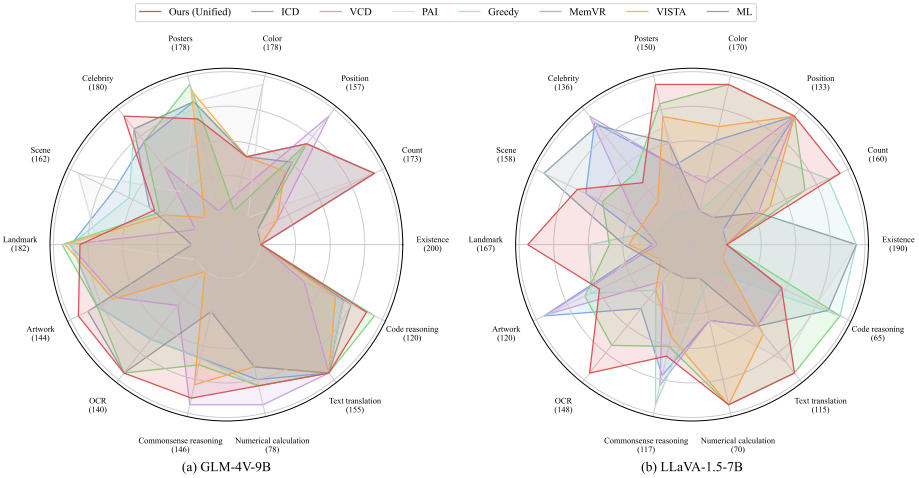


Figure 10: HallusionBench and LLaVA-Bench evaluation on Qwen2.5-VL-7B.

As shown in Fig. 10, VERGE achieves consistently strong performance on Qwen2.5-VL-7B across both HallusionBench and LLaVA-Bench, attaining the best results on most evaluation metrics. These results further demonstrate that VERGE effectively mitigates hallucinations while preserving strong general multimodal understanding and generation capabilities.



(a) GLM-4V-9B (b) LLaVA-1.5-7B

Figure 11: Reasoning Performance Across 14 MME Tasks.

Fig. 11 further presents the performance across 14 MME tasks. On GLM-4V-9B and LLaVA-1.5-7B, VERGE achieves the best or near-best performance on a large proportion of the evaluated tasks, with its radar profiles frequently forming or approaching the outer envelope among all compared methods. In particular, the gains are distributed across diverse dimensions, including object-level perception, OCR, numerical calculation, and text-related understanding, rather than being concentrated on a single capability. Although several individual tasks remain competitive with other methods, the overall radar coverage of VERGE is consistently broad. This indicates that the proposed hallucination mitigation does not compromise the general perception and reasoning capabilities of the backbone models, but instead maintains strong performance across a diverse set of multimodal tasks.

H.2 Ablation Study

We further conduct an ablation study to assess the contribution of the three components in VERGE. Specifically, S_1, S_2, and S_3 remove evidence-guided asymmetric channel correction, image-specific visual information reinforcement, and language-prior-calibrated contrastive decoding, respectively, while S_4 retains the complete framework. As shown in Tabel 5, across all three backbones, the full model consistently achieves the best performance, demonstrating that the three components contribute complementary benefits. Removing visual information reinforcement leads to clear degradation across models, highlighting the importance of recovering image-specific evidence when visual grounding weakens. Excluding contrastive decoding similarly reduces performance, confirming the value of calibrating residual language-prior dominance during generation. The consistent decline observed when channel correction is removed further indicates that rebalancing abnormal visual responses provides an important basis for faithful decoding. Overall, these results validate the coordinated design of VERGE and show that its effectiveness stems from the joint integration of channel correction, evidence recovery, and prior calibration.

Table 5: Ablation study on three models. ✓ and × indicate that the corresponding module is used or removed, respectively.

Scenario	Modules			Qwen2.5-VL-7B		InternVL3-8B		LLaVA-1.5-7B	
	rebalancing	rejection	CD	Score	Δ	Score	Δ	Score	Δ
S_1	×	✓	✓	2325.00	−3.50	2357.74	−0.25	1896.76	−1.76
S_2	✓	×	✓	2316.28	−12.22	2341.76	−16.23	1857.29	−41.23
S_3	✓	✓	×	2326.40	−2.10	2341.75	−16.24	1868.04	−30.48
S_4 **(Full)**	✓	✓	✓	**2328.50**	0.00	**2357.99**	0.00	**1898.52**	0.00

H.3 Parameter Analysis

Experimental setup. We analyze the sensitivity of VERGE to five key coefficients on Qwen2.5, InternVL3, LLaVA-1.5, and GLM-4V using the complete MME benchmark. Each configuration is evaluated on all 2,374 questions, comprising 1,187 image-shared question pairs across ten Perception and four Cognition categories. We report the full-MME score obtained by summing category-level accuracy and paired-question accuracy, with a maximum score of 2,800. This evaluation is distinct from the four-category MME-Hall setting used in the main experiments. The analyzed parameters include the evidence enhancement coefficient α, high-magnitude response suppression coefficient β, visual-token partition ratio ρ, visual reinforcement coefficient λ, and contrastive decoding coefficient δ. These coefficients govern the enhancement of evidence-aligned channels, attenuation of abnormal visual responses, construction of the high-magnitude token subset, incorporation of retrieved visual evidence, and calibration of the text-only prior, respectively. We vary one parameter at a time over the seven-point grids in Table 6, while keeping the remaining searched coefficients at the model-specific reference configuration. The reference tuples $(\alpha, \beta, \rho, \lambda, \delta)$ are $(0.6, 0.8, 0.1, 0.05, 0.05)$ for Qwen2.5, $(0.6, 0.8, 0.1, 0.05, 0.05)$ for GLM-4V, $(0.8, 1.2, 0.1, 0.001, 0.1)$ for InternVL3, and $(0.16, 0.25, 0.05, 0.001, 0.015)$ for LLaVA-1.5. The remaining model-specific settings are fixed to values tailored to each backbone. Specifically, $(L_s, L_e, c_{\min}, u_0)$ is set to $(6, 16, 0.2, 0.5)$ for Qwen2.5, $(5, 18, 0.2, 0.5)$ for GLM-4V, $(8, 12, 0.1, 0.75)$ for InternVL3, and $(5, 17, 0.8, 0.65)$ for LLaVA-1.5. Across all models, the prototype-pooling temperature is fixed at $\tau = 0.1$, while the plausibility schedule adopts $\kappa_0 = 0.1$, $\delta_{\text{ref}} = 1$, and $q = 2$. Other gate, image-processing, and decoding configurations remain unchanged within each backbone. This one-dimensional sweeping protocol isolates the local response to each coefficient while preserving a reproducible inference configuration.

Table 6: Selected seven-point, equally spaced search grids. Values are the nominal inputs used for the corresponding parameter sweeps.

Parameter	Qwen2.5 / GLM-4V	InternVL3	LLaVA-1.5
α	$\{0.4, 0.5, 0.6, 0.7, 0.8, 0.9, 1.0\}$	$\{0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2\}$	$\{0, 0.04, 0.08, 0.12, 0.16, 0.20, 0.24\}$
β	$\{0, 0.2, 0.4, 0.6, 0.8, 1.0, 1.2\}$	$\{0, 0.3, 0.6, 0.9, 1.2, 1.5, 1.8\}$	$\{0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$
ρ	$\{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35\}$	$\{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35\}$	$\{0.05, 0.10, 0.15, 0.20, 0.25, 0.30, 0.35\}$
λ	$\{0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$	$\{0, 0.001, 0.002, 0.003, 0.004, 0.005, 0.006\}$	$\{0, 0.0005, 0.0010, 0.0015, 0.0020, 0.0025, 0.0030\}$
δ	$\{0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$	$\{0, 0.05, 0.10, 0.15, 0.20, 0.25, 0.30\}$	$\{0, 0.005, 0.010, 0.015, 0.020, 0.025, 0.030\}$

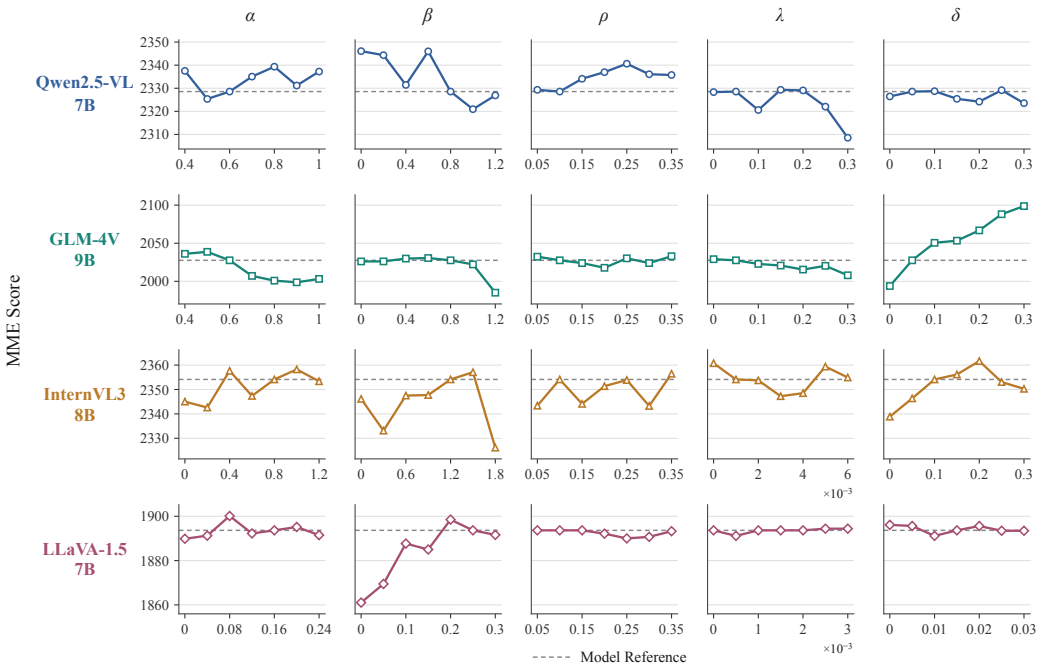


Figure 12: Parameter sensitivity across four MLLMs on the full MME benchmark. Each curve varies one coefficient while retaining the remaining settings at the corresponding model-specific reference configuration.

Table 7: Observed full-MME score spans computed as the difference between the maximum and minimum scores within each seven-point sweep. Bold values denote the largest span for each model.

Model	α	β	ρ	λ	δ
Qwen2.5	14.60	**58.03**	22.16	13.25	20.30
GLM-4V	45.32	75.25	16.36	18.05	**126.12**
InternVL3	15.63	**30.92**	13.22	13.53	22.81
LLaVA-1.5	10.29	**37.46**	3.64	3.25	4.97

Evidence enhancement: The coefficient α controls the strength of evidence-aligned channel enhancement. As shown in Fig. 12, its influence is clearly backbone-dependent and exhibits a nonmonotonic response across models. Qwen2.5 achieves its highest sampled score at $\alpha = 0.8$, while GLM-4V favors a moderate setting of $\alpha = 0.5$. InternVL3 attains its maximum score of 2358.23 at $\alpha = 1.0$, whereas LLaVA-1.5 reaches 1900.14 at $\alpha = 0.08$. These distinct operating regions indicate that the appropriate enhancement strength is closely coupled with the underlying representation characteristics of each backbone, supporting model-specific calibration rather than a universally shared setting.

High-magnitude response suppression: The coefficient β exhibits pronounced sensitivity across Qwen2.5, InternVL3, and LLaVA-1.5, as further reflected by the score spans in Table 7. LLaVA-1.5 improves from 1861.04 at $\beta = 0$ to 1898.50 at $\beta = 0.20$, while InternVL3 reaches its maximum at $\beta = 1.5$ before declining at the largest evaluated value. GLM-4V similarly favors an intermediate suppression strength of $\beta = 0.6$. For Qwen2.5, $\beta = 0$ and $\beta = 0.6$ yield nearly identical performance, differing by only 0.09 points. The nonmonotonic behavior across all four models shows that effective suppression requires balancing the removal of abnormal visual responses against the preservation of informative channel activity.

Visual-token partition: The partition ratio ρ determines the composition of the high-magnitude and complementary evidence subsets and therefore directly affects the visual prototypes used for channel correction. Qwen2.5 and GLM-4V obtain their highest scores at $\rho = 0.50$, while InternVL3 reaches its maximum at

$\rho = 0.35$. Since these values lie at the boundaries of the evaluated grids, they should be interpreted as favorable settings within the tested ranges rather than global optima. LLaVA-1.5 exhibits substantially weaker sensitivity, with a score span of only 3.64 points and identical aggregate scores at $\rho \in \{0.05, 0.10, 0.15\}$. The results suggest that the preferred partition granularity depends on the distribution of visual-token magnitudes induced by each backbone.

Visual reinforcement: The coefficient λ controls the contribution of retrieved image-specific evidence to the corrected FFN representation. Qwen2.5 and GLM-4V attain their highest scores at $\lambda = 0.20$ and $\lambda = 0.03$, respectively, while larger values eventually reduce performance. InternVL3 achieves its highest score at $\lambda = 0$, indicating that additional reinjection is unnecessary under the corresponding reference configuration. It also suggests that the remaining components already provide sufficiently strong hallucination mitigation in this regime, leaving limited room for further gains from visual reinjection. LLaVA-1.5 remains comparatively stable throughout its local search range, with a total span of 3.25 points. These observations show that visual reinforcement should complement rather than dominate the existing representation, with the appropriate mixing strength determined by the scale and structure of the backbone-specific hidden states.

Contrastive decoding: The coefficient δ controls the strength of language-prior calibration and produces the most pronounced response on GLM-4V. Its score increases throughout the evaluated range from 1993.85 at $\delta = 0$ to 2119.97 at $\delta = 0.40$. InternVL3 instead reaches its maximum of 2361.66 at $\delta = 0.20$ before declining at stronger settings. Qwen2.5 favors a weaker correction, reaching its highest score at $\delta = 0.03$, whereas LLaVA-1.5 performs best at the zero endpoint. Since δ simultaneously affects contrastive reranking and the adaptive plausibility threshold $\kappa(\delta) = \kappa_0 \min\{1, (\delta/\delta_{\text{ref}})^q\}$, these responses reflect the combined calibration of prior suppression and candidate plausibility. The results further demonstrate that the desirable degree of contrastive correction is highly dependent on the intrinsic prior structure of each backbone.

Overall analysis. Taken together, Fig. 12 and Table 7 reveal clear backbone-specific parameter responses rather than a shared sensitivity pattern across MLLMs. Suppression strength is particularly influential for Qwen2.5, InternVL3, and LLaVA-1.5, whereas contrastive decoding dominates the observed variation of GLM-4V. The remaining coefficients exhibit different degrees of local stability depending on the underlying representation and decoding characteristics. These results support model-specific parameter calibration while also providing empirical priors for practical deployment. In particular, the observed response ranges and stable operating regions can substantially narrow the search space when VERGE is transferred to related architectures, offering a useful reference for balancing hallucination mitigation and general multimodal capability.

H.4 VISUALIZATION OF HALLUCINATION MITIGATION

The following qualitative comparisons further illustrate that our method suppresses hallucinated content without compromising valid visual descriptions. In the examples shown, greedy decoding introduces unsupported objects, such as a parked car in COCO 136411 and a background tree in COCO 272953, whereas our method removes these spurious details while retaining the visually grounded content and overall semantic structure of the response. These cases suggest that the proposed approach improves visual grounding by selectively reducing unsupported generation rather than indiscriminately shortening or simplifying the output.

LLaVA-1.5-7B



COCO 136411

Greedy: The image features a stop sign with graffiti on it, located on a street corner. The graffiti consists of the words "Hammer Time" written underneath the word "Stop." The stop sign is positioned on a metal pole, and it is situated near a house. The scene also includes a car parked on the street, slightly behind the stop sign.

Ours: The image features a stop sign with graffiti written on it, making it appear vandalized. The graffiti reads "Hammer Time," which is a reference to the popular song by MC Hammer. The stop sign is located on a street corner, and there is a house visible in the background.



COCO 272953

Greedy: The image features a brown horse grazing on grass in a field. The horse is positioned in the center of the scene, with its head down as it eats the grass. The field is lush and green, providing an ideal environment for the horse to graze. In the background, there is a tree visible, adding to the natural setting of the scene. The horse appears to be enjoying its time in the field, peacefully grazing on the grass.

Ours: The image features a brown horse grazing on grass in a field. The horse is positioned in the center of the scene, with its head down as it eats the grass. The field is lush and green, providing an ideal environment for the horse to graze. The horse appears to be enjoying its time in the field, surrounded by the natural landscape.

Figure 13: Visualization of Hallucination Mitigation on LLava-1.5-7B.

LLaVA-1.5-13B



COCO 196118

Greedy: The image features a large clock tower, likely the Big Ben, towering over the city of London. The clock is prominently displayed on the side of the tower, making it a notable landmark. The tower is surrounded by a clear blue sky, which contrasts beautifully with the structure. In the foreground, there are two people standing near the base of the tower, possibly admiring the architecture or taking a photo. The scene captures the grandeur of the clock tower and the picturesque skyline of London.

Ours: The image features a large clock tower, which is a prominent landmark in the city. The clock tower is situated on top of a tall building, and its clock face is visible on the side of the tower. The clock is positioned towards the center of the tower, making it a focal point in the scene. The tower is surrounded by a blue sky, which adds to the overall beauty of the scene. The clock tower stands tall, towering over the city and serving as a symbol of the city's history and architectural heritage.



COCO 375755

Greedy: The image features a young boy standing on a sidewalk, holding a skateboard in his hands. He appears to be wearing a green shirt and is positioned near the center of the scene. The skateboard is placed vertically, with the boy holding it close to his body. In the background, there is a car parked on the street, and a bicycle can be seen further down the sidewalk. The scene captures the essence of a typical day in a neighborhood where children enjoy outdoor activities like skateboarding.

Ours: The image features a young boy standing on a sidewalk, holding a skateboard in his hands. He appears to be wearing a green shirt and is positioned near the center of the scene. The boy seems to be enjoying his time with the skateboard, possibly preparing to ride or just having finished a session. The skateboard is placed in his hands, and he appears to be the main focus of the image.

Figure 14: Visualization of Hallucination Mitigation on LLaVA-1.5-13B.

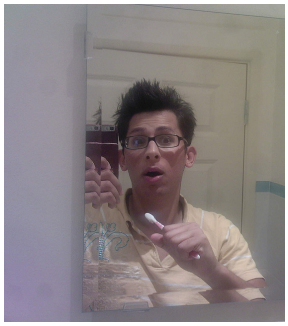
Qwen2.5-VL-7B



COCO 259824

Greedy: This image features an artistic and creative arrangement of vintage suitcases and luggage to form the shape of a chair. The chair is constructed with various suitcases of different sizes, colors, and materials, giving it a unique and eclectic appearance. - **Seat**: The seat of the chair is made from a large, dark brown suitcase that appears to be leather or a similar material. - **Backrest**: The backrest is formed by a green suitcase, which has a slightly worn texture, suggesting age and use. - **Armrests**: Two smaller suitcases, one black and one yellow, serve as armrests on either side of the chair. - **Legs**: The legs of the chair are also made from suitcases. The front legs consist of two stacked suitcases, one red and one brown, while the back legs are formed by a single brown suitcase. The chair is placed indoors, likely in a public space such as an airport or a waiting area, as suggested by the tiled floor and seating in the background. The lighting is bright, and the overall atmosphere is clean and modern. The juxtaposition of the old-fashioned suitcases with the contemporary setting creates a striking visual contrast.

Ours: This image features an artistic and creative arrangement of vintage suitcases and luggage to form the shape of a chair. The chair is constructed with various suitcases of different sizes, colors, and materials, giving it a unique and eclectic appearance. - **Backrest**: The backrest of the chair is made from a green suitcase, which appears slightly worn and has some visible scratches or marks on its surface. - **Seat**: The seat of the chair is formed by a dark brown suitcase, which looks sturdy and well-used. - **Armrests**: Two armrests are created using smaller suitcases. One is a yellow suitcase, and the other is a black suitcase, both of which have a more polished look compared to the others. - **Legs**: The legs of the chair are made from two rectangular suitcases, one red and one brown, which provide stability to the structure. The chair is placed indoors, likely in a public space such as an airport terminal or a waiting area, as suggested by the tiled floor, seating in the background, and the overall clean and open environment. The lighting is bright, indicating that the photo was taken during the day. The arrangement of the suitcases is not only functional but also serves as a conversation piece, showcasing creativity and resourcefulness.



COCO 488240

Greedy: The image shows a person standing in front of a mirror, brushing their teeth with a toothbrush in one hand and holding a smartphone in the other. The individual appears to be mid-expression, possibly reacting to something on the phone or simply posing for the photo. They are wearing glasses and a light-colored shirt with a pattern that includes blue and white elements. The background reveals part of a bathroom setting, including a door and what seems to be a shower curtain partially visible on the right side. The lighting is bright, likely from an overhead source, and the overall tone of the image suggests it was taken casually, perhaps during a morning routine.

Ours: The image shows a person standing in front of a mirror, brushing their teeth with a toothbrush in one hand and holding what appears to be a phone or a similar object in the other hand. The individual is wearing glasses and a light-colored shirt with a pattern that includes blue and beige tones. Their hair is styled upwards, giving it a spiky appearance. The background reflects part of a bathroom setting, including a door and some bathroom fixtures. The lighting in the room is bright, likely from an overhead source, which illuminates the scene clearly. The person's expression seems to convey surprise or excitement, as indicated by their wide eyes and open mouth.

Figure 15: Visualization of Hallucination Mitigation on Qwen2.5-VL-7B.

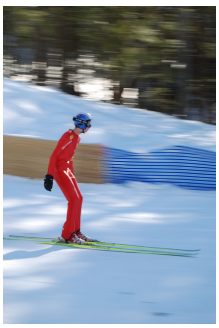
InternVL3-8B



COCO 122953

Greedy: The image shows a workspace setup with two computer monitors and a laptop. The laptop, positioned on the left, is open and displaying a Twitter page. The screen shows a typical Twitter interface with a text box and various tweets visible. The laptop is silver and has a trackpad below the keyboard. To the right is a larger desktop monitor, which also displays a similar Twitter page, along with several open windows, possibly for email or other other applications. The desktop monitor is older in style, with a bulky design. A cat is sitting between the laptop and the desktop monitor, partially in front of the desktop screen, looking to the side. In front of the monitors, there is a white keyboard and a mouse on a mousepad. The workspace appears to be on a desk with some cables and papers scattered around. The room has a window with blinds in the background.

Ours: The image shows a workspace setup with a laptop and an older model desktop monitor. The laptop is open and displaying a Twitter page, with a visible text box and some tweets. The desktop monitor next to it shows multiple open windows, including what appears to be a messaging or social media application and a spreadsheet or document. A cat is sitting between the laptop and the desktop monitor, partially in front of the desktop screen. The workspace has a keyboard in front of the laptop, and a mouse is visible on the right side. There are some cables and papers scattered on the desk. The room has a window with blinds in the background, allowing some natural light into the room. The overall setup gives a casual, home-office vibe.



COCO 563194

Greedy: The image shows a person skiing downhill. The skier is wearing a bright red ski suit and a blue helmet. They are in motion, as indicated by the blurred background, which suggests speed. The skis are green and yellow, and the skier is using poles. The background features a snowy landscape with trees and a blue and white striped barrier. The scene con to a dynamic and fast-paced skiing activity.

Ours: The image shows a person skiing downhill. The skier is wearing a bright red ski suit and a blue helmet. They are in motion, as indicated by the blurred background, which suggests speed. The skis are green and yellow, contrasting with the white snow. There are some trees and a blue and white striped barrier in the background, adding to the sense of movement and a possible ski race or slalom course. The scene is outdoors, likely in a snowy mountainous area.

Figure 16: Visualization of Hallucination Mitigation on InternVL3-8B.

GLM-4V-9B



COCO 438728

Greedy: This image captures a moment on a ski slope with a backdrop of a densely forested mountain. The foreground features two individuals engaged in winter sports. On the right, a person is dressed in a yellow and grey jacket, carrying a snowboard under their left arm, with ski poles in their right hand. They are wearing a dark blue beanie and a pair of goggles pushed up onto their head. The individual is walking on snow-covered ground, with a bright yellow chair visible in the lower right corner of the frame. In the background, another person is skiing down the slope. This skier is wearing a black and white jacket, grey pants, and is equipped with ski poles. The skier is in motion, indicated by the slight blur of their figure, and is wearing a pair of dark sunglasses pushed up onto their head. The slope is lined with trees, and in the distance, ski lifts can be seen transporting passengers up the mountain. The sky is clear, and the sunlight is casting shadows on the snow, indicating a sunny day. The overall scene is one of outdoor winter recreation and the enjoyment of a ski resort.

Ours: This image captures a moment on a ski slope with a backdrop of a densely forested mountain. The foreground is dominated by two individuals engaged in winter sports. On the right, there's a person dressed in a bright yellow and grey jacket, with a black beanie and ski goggles, walking with a snowboard under their arm. They appear to be in a hurry or perhaps taking a break from snowboarding. In the background, slightly blurred, is another individual who seems to be skiing. This person is wearing a black and white jacket, grey pants, and is equipped with ski poles. The skier is in motion, suggesting they are descending the slope. The slope is covered in snow, and the trees in the background are mostly evergreen, indicating a high altitude or a location that experiences heavy snowfall. The sky is clear, and the sun is shining, which suggests it's a day with good visibility and potentially good snow conditions for winter sports. In the distance, there are ski lift towers and chairs, which are common in ski resorts to transport visitors to the top of the slopes. The overall scene is one of outdoor activity and the enjoyment of winter sports in a beautiful mountainous landscape.



COCO 470924

Greedy: This image captures a group of four individuals seated at an outdoor dining table, enjoying a meal of pizza and beer. The table is set with plates of pizza, glasses of beer, and cutlery. The pizza appears to have various toppings, including what seems to be green bell peppers and onions. The group is dressed casually, with one person wearing a red jacket, another in a black hoodie, and the third in a black top. The fourth person is wearing a black top with a red trim. All four are wearing sunglasses, which could suggest it's a sunny day. The table is also adorned with a small bowl of what could be a side dish, and a silver napkin holder with a folded napkin. The background reveals an outdoor dining area with other patrons and tables, and a view of a busy street with cars and a white van parked along the side. The setting appears to be a relaxed, urban restaurant with a pleasant outdoor seating area.

Ours: This image captures a group of four individuals seated at a wooden table outdoors, enjoying a meal at what appears to be a restaurant with a view of a street. The table is laden with several pizzas of different toppings, and there are glasses of beer in front of each person. The setting is casual, with the patrons dressed in everyday clothing, and the presence of sunglasses suggests it might be a sunny day. The table also holds napkins, cutlery, and some empty plates, indicating that the meal is in progress or has just concluded. The background reveals a parking lot and a street bustling with vehicles, suggesting an urban location. The restaurant's outdoor seating area is furnished with wooden chairs and tables, and the ambiance is relaxed, with other patrons visible in the distance.

Figure 17: Visualization of Hallucination Mitigation on GLM-4V-9B.