BEYOND VISUAL ATTENTION: COMPETITION-INFORMED VISUAL EVIDENCE REFINEMENT FOR HALLUCINATION MITIGATION IN LVLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Hallucination remains a critical challenge for Large Vision-Language Models (LVLMs), where generated responses may be fluent yet inconsistent with the actual visual content. Existing studies commonly attribute hallucination to insufficient visual contribution and mitigate it by strengthening visual signals during decoding. However, hallucination may still occur even when the model attends strongly to the image. We find that hallucinated predictions exhibit greater overlap between the visual-evidence distributions associated with competing token candidates, suggesting that the attended visual information may not sufficiently distinguish these candidates. We term this phenomenon candidate-level visual-evidence ambiguity and find that it is associated with a higher rate of hallucination. Based on these findings, we propose Competition-Informed Visual Evidence Refinement (CIVER), a training-free decoding framework that disambiguates competing token predictions using four complementary components: a sample-specific visual prior, supporting evidence for the provisional leading candidate, competing evidence favoring its strongest alternative, and cross-step evidence memory. CIVER combines these components into a signed evidence bias and selectively applies it to subsequent visual attention according to the ambiguity of the candidate pair. Extensive experiments on POPE, CHAIR, MME, and AMBER using InstructBLIP, Shikra, LLaVA-1.5, and Qwen-VL demonstrate that CIVER consistently reduces hallucinations and improves visual faithfulness across diverse LVLM architectures.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) combine powerful language modeling capabilities with multimodal alignment mechanisms and have demonstrated remarkable performance in image captioning, visual question answering, and visual reasoning (Chen et al., 2023; Liu et al., 2023; Dai et al., 2023; Liu et al., 2024b). However, LVLMs remain susceptible to hallucinations and may generate content that is inconsistent with the visual input (Gunjal et al., 2024). For example, they may fabricate nonexistent objects or incorrectly describe visual attributes and relations, such as color, quantity, and spatial relationships (Gunjal et al., 2024; Li et al.; Huang et al.; Liu et al., 2024a). Such hallucinations substantially undermine the reliability of LVLMs, particularly in applications that require faithful visual understanding, including smart healthcare (Huang et al., 2025) and autonomous driving (Zhang et al., 2025).

To mitigate hallucination, some studies introduce additional alignment training (Yu et al., 2024b;a), or external supervision (Yin et al., 2023; Zhou et al., 2024). Although effective, these approaches often require substantial computational resources and task-specific data. In comparison, recent studies have increasingly focused on training-free methods applied during inference (Leng et al., 2024; Qiang et al., 2026; Yin et al., 2025; Mao et al., 2026; Wan et al., 2025; Lin et al., 2026; Wang et al., 2025; Xue et al.). Among them, contrastive decoding methods recalibrate next-token distributions by contrasting predictions obtained from the original image with those from perturbed or weakened visual inputs (Leng et al., 2024; Wang et al., 2024). Other methods directly intervene in internal representations or attention patterns to increase the contribution of visual information during generation (Qiang et al., 2026; Yin et al., 2025; Mao et al., 2026; Lin et al., 2026).

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.

Despite their different implementations, these approaches are largely motivated by the view that hallucination arises from insufficient visual contribution or excessive reliance on linguistic priors (Leng et al., 2024; Qiang et al., 2026; Yin et al., 2025; Wang et al., 2025; Chen et al., 2024). Accordingly, they generally rebalance decoding toward visual inputs by enhancing visual contributions or suppressing language-dominated predictions. However, these methods primarily operate on the overall amount of visual contribution, leaving unclear whether the attended visual information can reliably distinguish locally competing token candidates. This raises a more fine-grained question: does strong visual attention necessarily provide sufficiently discriminative evidence for next-token prediction?

We examine candidate-conditioned visual evidence during LVLM decoding. We conduct an event-level analysis of the LLaVA-1.5-7B baseline and obtain two key findings. First, hallucinations persist even when the model assigns relatively strong attention to image tokens, indicating that visual-attention strength alone does not characterize the quality of the attended information. Second, within both low- and high-attention groups, hallucination rates are higher when the visual-evidence distributions of the anchor-layer Top-2 candidates are less distinguishable. We term this phenomenon *candidate-level visual-evidence ambiguity*. These findings identify it as an important failure pattern associated with LVLM hallucination and suggest that uniformly strengthening visual signals may be insufficient when the current candidate competition remains unresolved.

Based on these findings, we propose **Competition-Informed Visual Evidence Refinement (CIVER)**, a training-free decoding framework for mitigating LVLM hallucinations. CIVER disambiguates competing token predictions using four complementary components: a sample-specific visual prior, supporting evidence for the provisional leading candidate, competing evidence favoring its strongest alternative, and cross-step evidence memory. It combines these components into a signed evidence bias and selectively applies the bias to evidence-aligned attention heads in subsequent decoder layers, with the intervention strength determined by the ambiguity of the candidate pair. In this way, CIVER improves candidate-level visual discrimination without uniformly amplifying visual signals or modifying model parameters.

Our main contributions are summarized as follows:

- We reveal two complementary patterns of LVLM hallucination: hallucinations can persist under relatively strong visual attention, and greater overlap between the visual-evidence distributions of competing token candidates is associated with higher hallucination rates. These findings identify candidate-level visual-evidence ambiguity as an important failure pattern associated with hallucination.
- We propose CIVER, a training-free decoding framework that compares the visual evidence associated with different token candidates and adjusts visual attention according to their relative support from the image. CIVER further uses an input-dependent visual prior and cross-step evidence memory as complementary evidence sources.
- Extensive experiments across four hallucination benchmarks and four representative LVLMs demonstrate that CIVER consistently reduces hallucinations while preserving general multimodal performance. Additional ablation and analysis experiments validate its main components and further support the identified failure mode.

2 Related Work

2.1 Training-Free Hallucination Mitigation in LVLMs.

Training-free methods mitigate hallucinations without updating model parameters and can be broadly divided into decoding calibration and internal representation intervention. Decoding-based approaches adjust next-token distributions by contrasting image-conditioned predictions with language-dominated or visually perturbed outputs. Intervention-based approaches instead strengthen visual representations or modify attention patterns within selected decoder layers. For example, ONLY (Wan et al., 2025) intervenes in text-biased attention heads, MemVR (Zou et al., 2025) reinjects visual tokens into intermediate layers under high uncertainty, and VAF (Yin et al., 2025) enhances visual signals in middle fusion layers. However, their interventions mainly regulate the overall influence of visual information without explicitly distinguishing the region-level

visual evidence associated with competing token candidates. Consequently, when multiple candidates receive similar visual support, strengthening the overall visual signal may reinforce their shared evidence without clarifying which candidate is better grounded in the image.

In contrast, CIVER compares candidate-conditioned visual evidence, models relative region-wise competition, and incorporates cross-step evidence memory to selectively redirect visual attention in subsequent decoder layers.

2.2 Visual Attention and Ambiguity Resolution

Visual information plays an important role in resolving ambiguous interpretations. In human vision, biased-competition theory suggests that concurrent representations compete for processing resources and that selective attention biases this competition toward decision-relevant evidence (Desimone et al., 1995). Recent multimodal studies similarly examine whether models can exploit visual context to resolve ambiguous language, revealing persistent difficulty in using image-specific cues consistently (Nam et al., 2025; Wang et al., 2026). However, these studies primarily address ambiguity in the input semantics rather than ambiguity in the visual evidence associated with competing decoder predictions. Fine-grained grounding methods establish correspondences between textual outputs and image regions, but stronger object grounding alone does not necessarily reduce hallucination (Geigle et al., 2024). These findings suggest that reliable generation requires not only access to visual information but also evidence that discriminates among competing token candidates. CIVER addresses this gap by comparing candidate-conditioned visual evidence and combining its relative support with an input-dependent visual prior and cross-step evidence memory to disambiguate candidate competition during decoding.

3 Preliminary Analysis and Motivation

Existing training-free methods commonly mitigate hallucination by increasing or restoring visual contribution during decoding. This view leaves open a more specific question: do hallucinated predictions always receive too little visual information, or can visual evidence be present yet insufficiently discriminative among competing token candidates? We investigate this question using the original LLaVA-1.5-7B model on CHAIR, with greedy decoding and a maximum generation length of 128 tokens. Object mentions are aligned with their corresponding decoding steps, and all statistics are computed over eligible single-token object events. The conditional hallucination rates reported below are therefore event-level quantities and differ from the standard caption-level CHAIR_S and object-level CHAIR_I metrics.

We estimate candidate-conditioned visual evidence at the intermediate anchor layer $n_a = 17$ using the same estimator later formalized in Eqs. 9–10. For candidates u and v at decoding step t, we quantify how clearly their visual support can be distinguished using the total variation distance

$$D_t(u, v) = \frac{1}{2} \left\| \mathbf{E}_t^u - \mathbf{E}_t^v \right\|_1 , \tag{1}$$

where $\mathbf{E}_t^u$ and $\mathbf{E}_t^v$ are normalized distributions over the image tokens. A smaller D_t means that the candidates rely on more overlapping estimated visual evidence and are therefore more ambiguous in terms of visual support.

We first examine the candidate evidence surrounding the model's actual final decision. Let y_t^{out} denote the token selected by the final layer and y_t^{alt} the highest-ranked anchor-layer candidate other than y_t^{out}. This final-output-centered pair is used only for retrospective diagnosis and does not constrain the candidates used by CIVER. After matching grounded and hallucinated events by decoding position, caption length, and anchor-layer rank, we retain 256 events in each group. Figure 1(a) shows that the mean evidence distance decreases from 0.0439 for grounded events to 0.0370 for hallucinated events. The difference remains statistically detectable under a Benjamini–Hochberg-corrected image-level permutation test ($p = 0.041$), with Cliff's $\delta = -0.276$. Hallucinated outputs are thus associated with greater visual-evidence overlap between the selected token and its strongest intermediate-layer alternative.

We next test whether the same failure pattern is observable from the native candidate pair available during inference. At each event, we take the anchor-layer Top-2 candidates (y_t^+, y_t^-) and compute

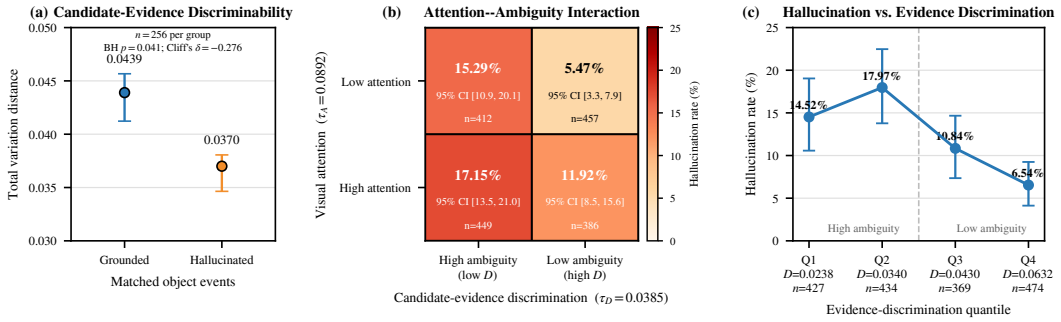


Figure 1: Preliminary analysis of candidate-level visual-evidence ambiguity on CHAIR using the original LLaVA-1.5-7B model. (a) Total variation distance between the evidence distributions of the final output token and its strongest anchor-layer alternative for matched grounded and hallucinated object events. (b) Conditional hallucination rates under combinations of visual-attention strength and native anchor-layer Top-2 evidence ambiguity. (c) Conditional hallucination rates across frozen evidence-discrimination quantile intervals for the native Top-2 pair. Lower D denotes more overlapping candidate evidence.

$D_t(y_t^+, y_t^-)$. We separately measure the total attention assigned to image tokens as

$$A_t = \frac{1}{|\mathcal{L}_a| H} \sum_{n \in \mathcal{L}_a} \sum_{h=1}^{H} \sum_{i \in \mathcal{I}} A_{t,i}^{(n,h)}, \tag{2}$$

where $\mathcal{I}$ is the set of image-token positions, H is the number of attention heads, and $\mathcal{L}_a = \{18, \ldots, 31\}$ is the subsequent-layer set. The median attention and evidence-distance thresholds are estimated exclusively on a calibration split and then frozen, giving $\tau_A = 0.0892$ and $\tau_D = 0.0385$. This yields 1,704 events on the held-out analysis split, with $D_t < \tau_D$ denoting high candidate ambiguity.

Figure 1(b) shows that candidate ambiguity remains associated with hallucination within both attention strata. Under low visual attention, the conditional hallucination rate rises from 5.47% for low-ambiguity events to 15.29% for high-ambiguity events. Under high visual attention, it similarly rises from 11.92% to 17.15%. In particular, the high-attention/high-ambiguity group has the highest hallucination rate among the four strata. Strong attention to image tokens therefore does not by itself imply that the attended evidence can distinguish the competing predictions; attention strength and candidate-evidence discrimination capture complementary properties of the decoding state.

Figure 1(c) verifies this relationship without relying on a single binary ambiguity threshold. Across the four frozen evidence-distance intervals, the conditional hallucination rates are 14.52%, 17.97%, 10.84%, and 6.54% as the median distance increases from 0.0238 to 0.0632. Although Q1 and Q2 are not strictly ordered, both higher-discrimination intervals exhibit lower hallucination rates than the two more ambiguous intervals. Together, the matched retrospective comparison and the inference-time Top-2 analysis identify candidate-level visual-evidence ambiguity as an important failure mode associated with hallucination. These findings motivate CIVER to compare the relative visual support of competing candidates and selectively refine subsequent visual attention, instead of relying only on a uniform increase in overall visual contribution. Consistent with this motivation, Appendix A presents a matched-prefix intervention analysis showing that CIVER improves original hallucinated-token suppression over VAF by 23.38 percentage points in the high-attention/high-ambiguity regime, with a 95% confidence interval of $[14.70, 32.47]$.

4 Method

4.1 Overview

We propose Competition-Informed Visual Evidence Refinement (CIVER), a training-free decoding method for mitigating hallucinations in large vision-language models. CIVER estimates and compares the visual evidence associated with the provisional leading candidate and its strongest alternative at an intermediate decoder layer, and uses their relative visual support to refine attention

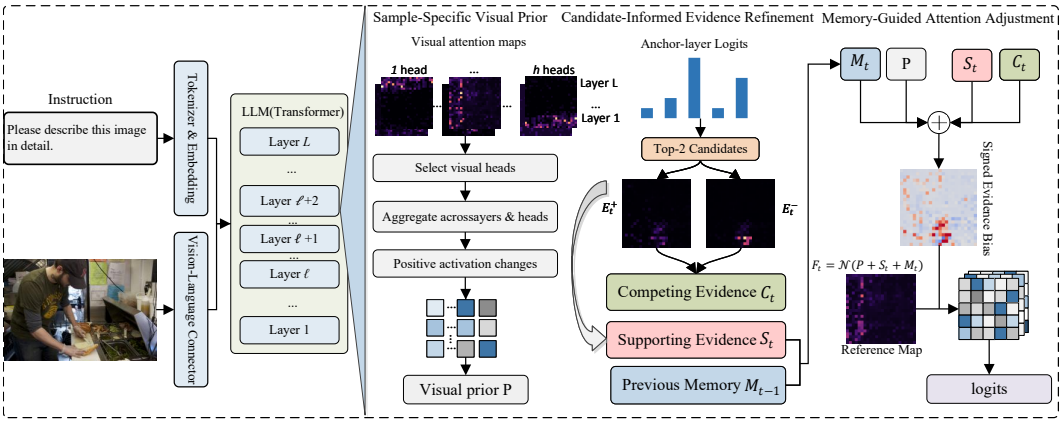


Figure 2: Overview of CIVER. CIVER first constructs a sample-specific visual prior from early decoder visual-attention patterns. At each decoding step, it extracts the top-2 provisional candidates at an intermediate anchor layer and compares their candidate-conditioned visual evidence to obtain supporting and competing evidence. The visual prior, current evidence, and cross-step memory are then integrated to construct a signed evidence bias and a reference map, which guide ambiguity-gated attention adjustment on evidence-aligned heads in subsequent decoder layers.

in subsequent layers. A sample-specific visual prior and cross-step evidence memory are further incorporated as complementary references.

Let the language decoder contain N Transformer layers. We define an intermediate anchor layer as

$$n_a = \lfloor \gamma_a N \rfloor , \tag{3}$$

where $\gamma_a \in [0, 1)$ controls its relative depth. Candidate tokens and their associated visual evidence are estimated at layer n_a, while attention adjustment is performed in the subsequent layers $\mathcal{L}_a = \{n_a + 1, \ldots, N - 1\}$. Thus, the anchor layer serves as an intermediate observation point rather than a source of definitive predictions.

For concise notation, $\mathcal{N}(\cdot)$, $\mathcal{Z}(\cdot)$, and $\overline{\mathbf{x}}$ denote nonnegative ℓ_1 normalization, standardization, and ℓ_2 normalization, respectively.

4.2 Sample-Specific Visual Prior

Evidence estimated at a single decoding step reflects the current prediction state and may be affected by the accompanying linguistic context. We therefore construct a sample-specific visual prior from the decoder's early visual-attention patterns to provide an input-level reference during subsequent generation.

CIVER first performs a greedy warm-up of T_w decoding steps. Let $A_{s,i}^{(n,h)}$ denote the attention assigned to visual token i at warm-up step s, layer n, and head h. For each step and layer, we retain the $K_v = \max(1, \lfloor H/2 \rfloor)$ heads with the largest total attention over the N_I visual tokens, denoted by $\mathcal{H}_s^{(n)}$. Their attention maps are averaged over the warm-up steps to obtain a layer-level visual map:

$$a_i^{(n)} = \frac{1}{T_w K_v} \sum_{s=1}^{T_w} \sum_{h \in \mathcal{H}_s^{(n)}} A_{s,i}^{(n,h)}. \tag{4}$$

A visual token is regarded as active at layer n when its attention exceeds a layer-wise threshold, $a_i^{(n)} > \tau_a \mu(\mathbf{a}^{(n)})$, and the corresponding binary indicator is denoted by $b_i^{(n)}$. We then accumulate the positive activation changes of each visual token across adjacent decoder layers:

$$d_i = \sum_{n=1}^{N-1} \left[b_i^{(n)} - b_i^{(n-1)} \right]_+ , \tag{5}$$

where $[x]_+ = \max(x, 0)$. A larger d_i indicates that visual token i becomes salient more frequently during information propagation across layers.

The accumulated scores are clipped, rescaled, and normalized to obtain the sample-specific visual prior:

$$\mathbf{P} = \mathcal{N}\left(\exp\left(\mathrm{MinMax}\left(\mathrm{Clip}_{3\sigma}(\mathbf{d})\right)\right)\right). \tag{6}$$

The resulting $\mathbf{P} \in \mathbb{R}^{N_I}$ provides a fixed reference over the visual tokens for the current image–instruction pair and remains unchanged during subsequent decoding.

4.3 Candidate-Informed Evidence Refinement

During multimodal prefill, we store the anchor-layer representations of all visual tokens, $\{\mathbf{v}_i^{(n_a)}\}_{i=1}^{N_I}$. At decoding step t, let $\mathbf{h}_t^{(n_a)}$ denote the hidden state of the current token position at the anchor layer. We apply the model's original final normalization and language-modeling head to obtain provisional vocabulary logits:

$$\widetilde{\mathbf{z}}_t = \mathbf{W}_U \,\mathrm{Norm}_f\left(\mathbf{h}_t^{(n_a)}\right), \tag{7}$$

where $\mathbf{W}_U$ is the output projection matrix.

We retain the two tokens with the largest provisional logits:

$$\left(y_t^+, y_t^-\right) = \mathrm{Top2}_{y\in\mathcal{V}}\, \widetilde{z}_t(y), \tag{8}$$

where y_t^+ denotes the provisional leading candidate and y_t^- its strongest alternative. The pair is used only to characterize the strongest local competition at the anchor layer and does not restrict the final prediction space. We further apply a pairwise softmax to their logits and denote the resulting relative probabilities by π_t^+ and π_t^-.

To estimate the visual evidence associated with each candidate, we jointly use the current decoding state and the corresponding output vector. Let $\mathbf{e}_t^k = \mathbf{W}_U[y_t^k, :]^\top$, where $k \in \{+, -\}$. We construct the candidate-conditioned query as

$$\mathbf{q}_t^k = \mathrm{L2Norm}\left((1-\eta)\overline{\mathbf{h}_t^{(n_a)}} + \eta\overline{\mathbf{e}_t^k}\right), \tag{9}$$

where $\eta \in [0, 1]$ balances contextual and candidate-specific information.

The compatibility between the query and each anchor-layer visual-token representation is then normalized over the visual tokens:

$$E_{t,i}^k = \frac{\exp\left((\mathbf{q}_t^k)^\top \overline{\mathbf{v}_i^{(n_a)}}/\tau_e\right)}{\sum_{j=1}^{N_I}\exp\left((\mathbf{q}_t^k)^\top \overline{\mathbf{v}_j^{(n_a)}}/\tau_e\right)}, \tag{10}$$

where τ_e is the evidence temperature. The resulting distributions $\mathbf{E}_t^+$ and $\mathbf{E}_t^-$ describe the relative visual evidence associated with the two provisional candidates.

Because the two candidates may rely on overlapping visual regions, directly suppressing the entire evidence distribution of the alternative candidate could also weaken regions shared with the leading candidate. We therefore use $\mathbf{S}_t = \mathbf{E}_t^+$ as the evidence associated with the current leading candidate and retain only the relative excess of the alternative candidate:

$$\mathbf{C}_t = \begin{cases} \mathcal{N}\left([\mathbf{E}_t^- - \mathbf{E}_t^+]_+\right), & \left\|[\mathbf{E}_t^- - \mathbf{E}_t^+]_+\right\|_1 > \epsilon, \\ \mathbf{0}, & \text{otherwise.} \end{cases} \tag{11}$$

Thus, $\mathbf{C}_t$ removes candidate-shared evidence and retains only visual regions that comparatively favor y_t^-.

The strength of the current candidate competition is characterized using the pairwise probability margin $m_t = \pi_t^+ - \pi_t^-$ and evidence overlap $o_t = \sum_i \min(E_{t,i}^+, E_{t,i}^-)$. A small margin or large overlap indicates a more ambiguous prediction state. We combine these two factors into an adaptive gate:

$$g_t = \sigma\left(\frac{\lambda_m(1-m_t) + \lambda_o o_t - \delta_g}{\tau_g}\right), \tag{12}$$

where λ_m and λ_o control the contributions of the probability margin and evidence overlap, δ_g is the activation threshold, and τ_g controls the smoothness of the gate. A larger g_t therefore corresponds to a less clearly separated candidate pair.

4.4 Memory-Guided Attention Adjustment

The evidence estimated at an individual decoding step provides only a local view of the current prediction state. To introduce a complementary cross-step reference, CIVER maintains an exponentially averaged evidence memory:

$$\mathbf{M}_t = \begin{cases} \mathbf{S}_t, & t = 1, \\ \mathcal{N}\left(\rho \mathbf{M}_{t-1} + (1-\rho)\mathbf{S}_t\right), & t > 1, \end{cases} \tag{13}$$

where $\rho \in [0, 1)$ is the memory-decay coefficient. The memory summarizes the supporting evidence accumulated across the current and previous decoding steps rather than imposing historical evidence as a fixed constraint.

We then integrate the sample-specific prior, current candidate evidence, competing evidence, and cross-step memory into a signed evidence bias:

$$\mathbf{B}_t = \mathrm{Clip}\Big(\alpha_p \mathcal{Z}(\mathbf{P}) + \alpha_s \mathcal{Z}(\mathbf{S}_t) - \alpha_c \mathcal{Z}(\mathbf{C}_t) + \alpha_m \mathcal{Z}(\mathbf{M}_t), -b_{\max}, b_{\max}\Big). \tag{14}$$

where α_p, α_s, α_c, and α_m control the contributions of the four evidence sources. Since each map is standardized before combination, positive entries of $\mathbf{B}_t$ increase the corresponding visual-token logits, whereas negative entries decrease them.

Applying the same adjustment to all attention heads may interfere with heads that are weakly related to the current visual evidence. We therefore construct a reference map

$$\mathbf{F}_t = \mathcal{N}\left(\mathbf{P} + \mathbf{S}_t + \mathbf{M}_t\right)$$

and measure its agreement with the within-image attention distribution of each head:

$$\kappa_t^{(n,h)} = \sum_{i=1}^{N_I} \mathrm{Softmax}\left(\mathbf{L}_{t,I}^{(n,h)}\right)_i F_{t,i}, \tag{15}$$

where $\mathbf{L}_{t,I}^{(n,h)}$ denotes the pre-softmax attention logits over visual tokens at layer n and head h. For each layer $n \in \mathcal{L}_a$, we retain the $K_h = \max(1, \mathrm{round}(\gamma_h H))$ heads with the largest agreement scores and denote the selected set by $\mathcal{H}_t^{(n)}$.

Finally, the ambiguity-gated evidence bias is added only to the visual-token logits of the selected heads:

$$\widetilde{\mathbf{L}}_{t,I}^{(n,h)} = \begin{cases} \mathbf{L}_{t,I}^{(n,h)} + g_t \mathbf{B}_t, & h \in \mathcal{H}_t^{(n)}, \\ \mathbf{L}_{t,I}^{(n,h)}, & h \notin \mathcal{H}_t^{(n)}. \end{cases} \tag{16}$$

The adjusted logits are passed through the model's original attention softmax, after which the remaining forward computation proceeds unchanged. CIVER therefore operates entirely at inference time: it does not modify the vision encoder, update model parameters, or restrict the final vocabulary. The full algorithm of CIVER is provided in Appendix B.

5 Experiments

5.1 Experiment Settings

Backbones. We evaluate four representative 7B-scale LVLMs: InstructBLIP (Dai et al., 2023), Shikra (Chen et al., 2023), LLaVA-1.5 (Liu et al., 2024b), and Qwen-VL (Bai et al., 2023). InstructBLIP compresses visual features through a Q-Former, Shikra extends LLaVA with spatial grounding capabilities, LLaVA-1.5 aligns visual and language features using a projection-based connector, and Qwen-VL performs multimodal fusion through cross-attention. These models cover diverse vision-language architectures and fusion strategies.

Baselines. We compare CIVER with three recent training-free hallucination-mitigation methods. VAF (Yin et al., 2025) strengthens visual signals in middle fusion layers to improve visual sensitivity while preserving generation quality and efficiency. HulluEdit (Lin et al., 2026) performs single-pass orthogonal subspace editing to suppress conflicting priors while preserving visual evidence. PM (Mao et al., 2026) iteratively identifies attention-relevant visual regions and magnifies them to enhance fine-grained visual perception during decoding.

Datasets. (1) POPE (Li et al.) evaluates object hallucination using balanced Yes/No questions from MSCOCO (Lin et al.), A-OKVQA (Schwenk et al.), and GQA (Hudson & Manning). It adopts random, popular, and adversarial sampling for absent objects and contains 27,000 question-answer pairs. We report accuracy and F1. (2) CHAIR (Rohrbach et al.) evaluates object hallucination in generated captions. CHAIR_I measures the proportion of hallucinated objects, while CHAIR_S measures the proportion of captions containing hallucinations. We evaluate 500 MSCOCO validation images using maximum generation lengths of 64 and 128 tokens. (3) MME (Fu et al., 2026) contains ten perceptual and four cognitive subtasks. Following (Yin et al., 2023), we report the full score and four hallucination-related subsets: existence, counting, position, and color. (4) AMBER (Wang et al., 2023) evaluates existence, attribute, and relation hallucinations in generative and discriminative settings using 1,004 annotated images. We report CHAIR, Cover, HAL, and COG, where lower CHAIR, HAL, and COG and higher Cover indicate better performance.

5.2 Experiment Results

Results on POPE Table 1 reports the average POPE results over MSCOCO, A-OKVQA, and GQA. CIVER consistently achieves the best Accuracy and F1 scores across all four LVLMs under the Random, Popular, and Adversarial settings. Compared with the strongest competing method for each model and metric, CIVER yields average improvements of 0.95, 1.51, and 1.46 points under the three settings, respectively, with an overall average gain of 1.31 points across all 24 comparisons.

The improvements are more pronounced under the more challenging Popular and Adversarial settings. Under Popular sampling, CIVER improves F1 by 1.76, 2.26, 1.56, and 1.56 points on InstructBLIP, Shikra, LLaVA-1.5, and Qwen-VL, respectively. Under Adversarial sampling, it increases Accuracy by 2.00 points on InstructBLIP and 2.33 points on Qwen-VL, while improving F1 by 1.54 points on Shikra and 1.87 points on Qwen-VL.

Table 1: Average POPE results across MSCOCO, A-OKVQA, and GQA under the Random, Popular, and Adversarial settings. Best results are highlighted in bold.

Setting	Decoding	InstructBLIP		Shikra		LLaVA-1.5		Qwen-VL	
		Acc	F1	Acc	F1	Acc	F1	Acc	F1
Random	Regular	80.23	80.88	80.58	81.05	83.16	82.42	85.00	84.08
	VAF	86.41	86.75	85.53	86.11	87.94	88.40	88.87	89.03
	HulluEdit	86.85	87.75	84.93	85.20	88.25	88.05	89.27	89.85
	PM	86.71	87.63	85.37	86.02	88.45	88.14	89.13	89.29
	CIVER	**87.47**	**88.61**	**86.35**	**87.50**	**89.98**	**88.91**	**90.26**	**90.75**
Popular	Regular	75.31	77.01	75.14	75.82	79.20	79.15	82.03	81.22
	VAF	80.34	80.85	79.64	80.09	81.12	82.81	86.23	86.22
	HulluEdit	80.95	80.20	79.51	79.34	82.61	81.82	87.02	87.27
	PM	79.39	79.85	80.48	80.26	81.67	81.44	86.75	86.87
	CIVER	**81.85**	**82.61**	**82.35**	**82.52**	**83.04**	**84.37**	**88.79**	**88.83**
Adversarial	Regular	70.85	73.00	70.93	72.35	71.75	75.02	79.66	79.19
	VAF	74.97	77.19	74.79	75.39	74.51	78.13	83.93	83.74
	HulluEdit	75.05	75.71	75.62	76.03	76.01	77.24	84.20	84.35
	PM	74.69	75.08	75.14	75.31	75.24	73.97	84.25	85.01
	CIVER	**77.05**	**77.34**	**77.27**	**77.57**	**76.84**	**79.45**	**86.58**	**86.88**

Results on CHAIR Tables 2 and 3 report the CHAIR results with maximum generation lengths of 64 and 128 tokens, respectively. CIVER consistently achieves the lowest C_S and C_I scores across all four LVLMs. Under the 64-token setting, it reduces C_S and C_I by an average of 4.15 and 1.73 points, respectively, compared with the strongest competing baselines. The advantage becomes more pronounced for 128-token generation, where the average reductions increase to 8.25 points in C_S and 3.15 points in C_I. These results demonstrate that CIVER effectively suppresses both sentence-level and instance-level object hallucinations, with particularly strong improvements

Table 2: Results on CHAIR (max new tokens = 64).

	InstructBLIP		Shikra		LLaVA-1.5		Qwen-VL	
	C_S	C_I	C_S	C_I	C_S	C_I	C_S	C_I
Regular	25.8	9.7	24.7	8.7	22.0	6.4	24.1	7.8
VAF	22.6	6.7	22.8	7.5	22.2	6.8	20.9	6.7
HulluEdit	19.7	6.7	21.4	7.5	19.2	6.5	18.9	6.4
PM	20.1	7.3	20.8	7.6	22.0	8.1	19.7	6.9
Ours	**16.1**	**5.1**	**16.8**	**5.4**	**14.3**	**4.7**	**14.8**	**4.9**

Table 3: Results on CHAIR (max new tokens = 128).

	InstructBLIP		Shikra		LLaVA-1.5		Qwen-VL	
	C_S	C_I	C_S	C_I	C_S	C_I	C_S	C_I
Regular	55.6	17.6	54.3	15.7	49.6	13.8	51.8	16.7
VAF	50.8	14.5	48.5	14.1	48.4	13.5	44.5	13.1
HulluEdit	48.2	14.7	50.1	14.8	47.0	14.3	45.9	13.6
PM	49.7	15.6	47.2	14.3	46.2	15.3	47.5	14.2
Ours	**40.2**	**11.5**	**39.0**	**11.2**	**37.8**	**9.7**	**36.1**	**10.2**

during longer generation. This trend is consistent with its cross-step evidence refinement, which provides complementary visual evidence throughout the decoding process.

Results on MME Fig. 3 presents the results on the 14 MME subtasks across four LVLMs. CIVER achieves overall scores of 1992.31, 1551.08, 1682.74, and 2031.13 on LLaVA-1.5, Shikra, InstructBLIP, and Qwen-VL, respectively, outperforming the strongest competing methods by an average of 109.99 points. The improvements are particularly evident on hallucination-related perceptual tasks such as existence, counting, position, and color, while CIVER also maintains strong performance on knowledge- and reasoning-oriented dimensions. The broadly expanded radar curves across different backbones indicate that CIVER improves visual faithfulness without sacrificing general multimodal capabilities.

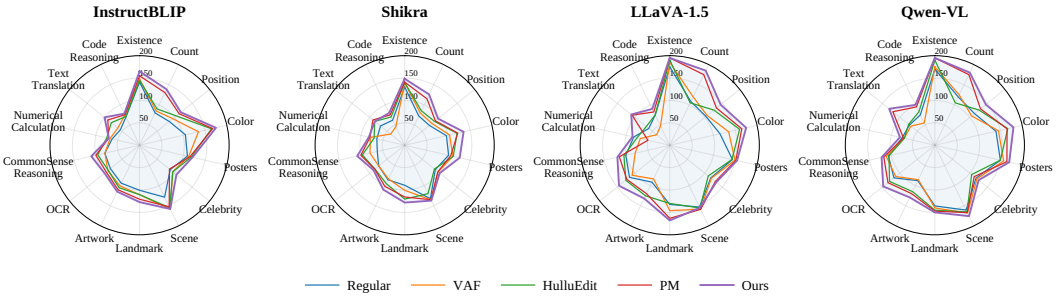


Figure 3: Comparison on the 14 MME subtasks across different LVLMs.

6 Conclusion

We proposed CIVER, a training-free decoding framework for mitigating object hallucination in LVLMs. Instead of uniformly strengthening visual information, CIVER focuses on whether the available visual evidence can sufficiently distinguish competing token candidates. To this end, it integrates a sample-specific visual prior, candidate-conditioned supporting and competing evidence, and cross-step memory, and uses the resulting evidence signals to selectively refine visual attention during decoding. Experiments on POPE, CHAIR, MME, and AMBER show that CIVER consistently improves hallucination-related performance across multiple LVLM backbones, with particularly clear gains under challenging evaluation settings and longer generation. These results suggest that modeling relative visual support among competing candidates is a promising direction for improving visual faithfulness without requiring additional training or parameter updates.

AI use statement

Generative AI tools were used to assist with language editing and polishing of the manuscript, including improving grammar, clarity, and readability. All AI-assisted edits were carefully reviewed and verified by the authors. The authors take full responsibility for the final content of the paper.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *CoRR*, abs/2308.12966, 2023. doi: 10.48550/ARXIV.2308.12966. URL https://doi.org/10.48550/arXiv.2308.12966.

Keqin Chen, Zhao Zhang, Weili Zeng, et al. Shikra: Unleashing multimodal llm's referential dialogue magic. *CoRR*, abs/2306.15195, 2023. doi: 10.48550/ARXIV.2306.15195.

Zhaorun Chen, Zhuokai Zhao, et al. HALC: object hallucination reduction via adaptive focal-contrast decoding. In *Forty-first International Conference on Machine Learning, ICML 2024*, 2024.

Wenliang Dai, Junnan Li, Dongxu Li, et al. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *Advances in Neural Information Processing Systems 36: NeurIPS 2023*, 2023.

Robert Desimone, John Duncan, et al. Neural mechanisms of selective visual attention. *Annual review of neuroscience*, 18(1):193–222, 1995.

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.

Gregor Geigle, Radu Timofte, and Goran Glavaš. Does object grounding really reduce hallucination of large vision-language models? In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 2728–2742, 2024.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Thirty-Eighth AAAI Conference on Artificial Intelligence, AAAI 2024*, pp. 18135–18143, 2024. doi: 10.1609/AAAI.V38I16.29771.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. Visual hallucinations of multi-modal large language models. In *Findings of the Association for Computational Linguistics, ACL 2024*, pp. 9614–9631. doi: 10.18653/V1/2024.FINDINGS-ACL.573.

Xijie Huang, Xinyuan Wang, Hantao Zhang, Yinghao Zhu, Jiawen Xi, Jingkun An, Hao Wang, Hao Liang, and Chengwei Pan. Medical mllm is vulnerable: cross-modality jailbreak and mismatched attacks on medical multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 3797–3805, 2025.

Drew A. Hudson and Christopher D. Manning. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019,*, pp. 6700–6709. doi: 10.1109/CVPR.2019.00686.

Sicong Leng, Hang Zhang, et al. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pp. 13872–13882, 2024. doi: 10.1109/CVPR52733.2024.01316.

Yifan Li, Yifan Du, et al. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023*, pp. 292–305. doi: 10.18653/V1/2023.EMNLP-MAIN.20.

Tsung-Yi Lin, Michael Maire, et al. Microsoft COCO: common objects in context. In *ECCV 2014*, volume 8693, pp. 740–755. doi: 10.1007/978-3-319-10602-1_48.

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

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. In *Advances in Neural Information Processing Systems 36, NeurIPS 2023*, 2023.

Haotian Liu, Chunyuan Li, et al. Improved baselines with visual instruction tuning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pp. 26286–26296, 2024b. doi: 10.1109/CVPR52733.2024.02484.

Shunqi Mao, Chaoyi Zhang, and Weidong Cai. Through the magnifying glass: Adaptive perception magnification for hallucination-free vlm decoding. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 44480–44501, 2026.

Heejeong Nam, Jinwoo Ahn, Keummin Ka, Jiwan Chung, and Youngjae Yu. Vague: Visual contexts clarify ambiguous expressions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 1537–1547, October 2025.

Pengpeng Qiang, Hongye Tan, Hu Zhang, Xiaoli Li, Ru Li, and Jiye Liang. Mitigating hallucinations in large vision-language models via visual-enhanced contrastive decoding. *IEEE Trans. Multim.*, 28:3242–3255, 2026. doi: 10.1109/TMM.2026.3651099.

Anna Rohrbach, Lisa Anne Hendricks, et al. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,EMNLP2018*, pp. 4035–4045. doi: 10.18653/V1/D18-1437.

Dustin Schwenk, Apoorv Khandelwal, et al. A-OKVQA: A benchmark for visual question answering using world knowledge. In *ECCV 2022*, volume 13668, pp. 146–162. doi: 10.1007/978-3-031-20074-8_9.

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

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. MLLM can see? dynamic correction decoding for hallucination mitigation. In *The Thirteenth International Conference on Learning Representations, ICLR 2025, Singapore, April 24-28, 2025*. OpenReview.net, 2025. URL https://openreview.net/forum?id=4z3IguA4Zg.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *CoRR*, abs/2311.07397, 2023. doi: 10.48550/ARXIV.2311.07397. URL https://doi.org/10.48550/arXiv.2311.07397.

Ru Wang, Selena Song, Yuquan Wang, Liang Ding, Mingming Gong, Yusuke Iwasawa, Yutaka Matsuo, and Jiaxian Guo. Mma:benchmarking multi-modallarge language models in ambiguity contexts. In Rebekka Burkholz, Shiwei Liu, Saiprasad Ravishankar, William Redman, Wei Huang, Weijie Su, and Zhihui Zhu (eds.), *Conference on Parsimony and Learning*, volume 328 of *Proceedings of Machine Learning Research*, pp. 529–551. PMLR, 23–26 Mar 2026.

Xintong Wang, Jingheng Pan, et al. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics, ACL 2024*, pp. 15840–15853, 2024. doi: 10.18653/V1/2024.FINDINGS-ACL.937.

Hongwei Xue, Yupan Huang, Bei Liu, et al. Probing inter-modality: Visual parsing with self-attention for vision-and-language pre-training. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021*, pp. 4514–4528.

Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference (CVPR)*, pp. 14625–14634, June 2025.

Shukang Yin, Chaoyou Fu, et al. Woodpecker: Hallucination correction for multimodal large language models. *CoRR*, abs/2310.16045, 2023. doi: 10.48550/ARXIV.2310.16045. URL https://doi.org/10.48550/arXiv.2310.16045.

Qifan Yu, Juncheng Li, et al. Hallucidoctor: Mitigating hallucinatory toxicity in visual instruction data. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR*, pp. 12944–12953, 2024a. doi: 10.1109/CVPR52733.2024.01230.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, and Maosong Sun. RLHF-V: towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024*, pp. 13807–13816, 2024b.

Ruifei Zhang, Wei Zhang, Xiao Tan, Sibei Yang, Xiang Wan, Xiaonan Luo, and Guanbin Li. Vldrive: Vision-augmented lightweight mllms for efficient language-grounded autonomous driving. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 5923–5933, 2025.

Yiyang Zhou, Chenhang Cui, et al. Analyzing and mitigating object hallucination in large vision-language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024*, 2024.

Xin Zou, Yizhou Wang, Yibo Yan, et al. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. In *Forty-second International Conference on Machine Learning*, 2025.

A Intervention Analysis under Candidate Ambiguity

Section 3 shows that candidate-evidence ambiguity is associated with hallucination even when the model assigns relatively strong attention to image tokens. We complement that observational analysis with a matched-prefix intervention experiment designed to test whether CIVER is particularly effective in this failure regime. We consider the 211 originally hallucinated single-token object events from the held-out CHAIR analysis split. Each event is assigned to one of the four attention–ambiguity strata using the original Regular trajectory and the frozen thresholds $\tau_A = 0.0892$ and $\tau_D = 0.0385$. The assignments are fixed before intervention and are not recomputed for either VAF or CIVER.

For every event, Regular, VAF, and CIVER are evaluated from the same decoding prefix. Let y_e^{reg} be the originally generated hallucinated token and $\widehat{y}_e^{m}$ the token produced by method m from that prefix. We define original hallucinated-token suppression as

$$S_m(\mathcal{G}) = \frac{100}{|\mathcal{G}|} \sum_{e \in \mathcal{G}} \mathbb{I}\left[\widehat{y}_e^{m} \neq y_e^{\text{reg}}\right], \tag{17}$$

where $\mathcal{G}$ denotes a frozen stratum. This metric measures whether an intervention dislodges the original hallucinated token; it does not by itself require the replacement token to be visually grounded. Pairwise gains are computed event by event, and confidence intervals are obtained using image-cluster bootstrap so that events from the same image remain grouped.

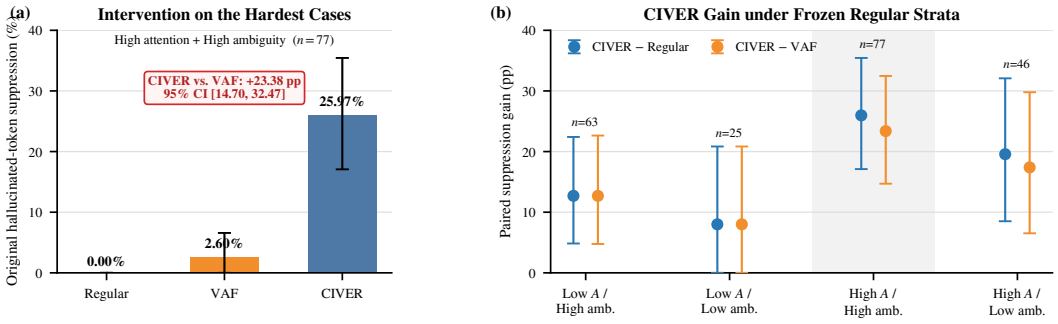


Figure 4: Matched-prefix intervention analysis on originally hallucinated CHAIR object events. Attention–ambiguity strata are assigned once using the original Regular trajectory and remain fixed for all methods. (a) Suppression of the original hallucinated token by Regular, VAF, and CIVER in the high-attention/high-ambiguity stratum. Regular is zero by construction under deterministic decoding. (b) Paired suppression gains of CIVER over Regular and VAF in all four frozen strata; positive values favor CIVER. Error bars denote 95% image-cluster bootstrap confidence intervals.

Figure 4(a) focuses on the high-attention/high-ambiguity stratum identified in Section 3 as the clearest counterexample to an attention-deficiency-only explanation. Among its 77 originally hallucinated events, VAF suppresses only 2.60% of the original hallucinated tokens, whereas CIVER suppresses 25.97%. The paired CIVER–VAF gain is $+23.38$ percentage points, with a 95% image-cluster bootstrap confidence interval of $[14.70, 32.47]$. Thus, when visual attention is already high but candidate-specific evidence remains strongly overlapping, candidate-informed refinement is substantially more effective than further visual amplification.

The full stratified comparison in Fig. 4(b) shows that CIVER suppresses 12.70%, 8.00%, 25.97%, and 19.57% of the original hallucinated tokens in the low-attention/high-ambiguity, low-attention/low-ambiguity, high-attention/high-ambiguity, and high-attention/low-ambiguity strata, respectively. Its paired gains over VAF are 12.70, 8.00, 23.38, and 17.39 percentage points in the same order. The intervals are relatively wide in the smaller strata, especially for the low-attention/low-ambiguity group with only 25 events, so differences among all four strata should be interpreted descriptively. Nevertheless, the largest and most clearly estimated advantage occurs in the high-attention/high-ambiguity regime. This result is consistent with the design of CIVER: the method uses candidate-relative visual support to redirect attention when visual information is present but does not clearly resolve the current candidate competition.

B ALGORITHM

Algorithm 1 summarizes the complete decoding procedure of CIVER.

Algorithm 1 Competition-Informed Visual Evidence Refinement (CIVER)

Require: Image I, prompt X, LVLM f_θ, maximum length $T_{\max}$
Ensure: Generated sequence Y

$Y \leftarrow \emptyset$, $n_a \leftarrow \lfloor \gamma_a N \rfloor$
Construct sample-specific visual prior $\mathbf{P}$
Store anchor-layer visual representations $\{\mathbf{v}_i^{(n_a)}\}_{i=1}^{N_I}$
for $t = 1, \ldots, T_{\max}$ **do**
  Obtain anchor hidden state $\mathbf{h}_t^{(n_a)}$
  Compute provisional logits $\widetilde{\mathbf{z}}_t$
  $(y_t^+, y_t^-) \leftarrow \mathrm{Top2}(\widetilde{\mathbf{z}}_t)$
  Compute candidate evidence $\mathbf{E}_t^+$ and $\mathbf{E}_t^-$
  Compute supporting evidence $\mathbf{S}_t$ and competing evidence $\mathbf{C}_t$
  Compute ambiguity gate g_t
  Update cross-step memory $\mathbf{M}_t$
  Construct signed evidence bias $\mathbf{B}_t$
  for $n = n_a + 1, \ldots, N-1$ **do**
    Select evidence-aligned heads $\mathcal{H}_t^{(n)}$
    for $h \in \mathcal{H}_t^{(n)}$ **do**
      $\widetilde{\mathbf{L}}_{t,I}^{(n,h)} \leftarrow \mathbf{L}_{t,I}^{(n,h)} + g_t \mathbf{B}_t$
    end for
  end for
  Decode y_t and append it to Y
  if y_t is EOS **then**
    break
  end if
end for
return Y

C IMPLEMENTATION DETAILS

We use a unified configuration across all models and datasets, with the anchor-layer ratio, visual-prior scale, memory-decay coefficient, and selected-head ratio set to $\gamma_a = 0.55$, $\alpha_p = 0.8$, $\rho = 0.8$, and $\gamma_h = 0.5$, respectively. For an LVLM with N decoder layers, the anchor layer is $n_a = \lfloor \gamma_a N \rfloor$, and attention refinement is applied to the subsequent layers $\mathcal{L}_a = \{n_a + 1, \ldots, N-1\}$. At each decoding step, the top two anchor-layer candidates are used to characterize local candidate competition, while all evidence-dependent quantities are determined dynamically. All experiments are conducted on a single NVIDIA 5880ADA GPU.

D ADDITIONAL EXPERIMENTS

Fig. 5 presents the AMBER results across four LVLM backbones. CIVER achieves the best performance on all four metrics for InstructBLIP, Shikra, LLaVA-1.5, and Qwen-VL, yielding the best result in all 16 model–metric comparisons. Compared with the strongest competing baseline, CIVER reduces CHAIR, Hal, and Cog by an average of 0.90, 3.53, and 0.50 points, respectively, while improving Cover by 3.43 points on average.

The gains are particularly notable on Qwen-VL, where CIVER increases Cover from 64.9 to 72.7 and reduces Hal from 46.2 to 41.0. Similar improvements are observed across the other backbones, indicating that CIVER not only suppresses object hallucinations but also preserves a higher proportion of visually grounded content. The consistently expanded radar profiles further demonstrate that the effectiveness of CIVER generalizes across different LVLM architectures.

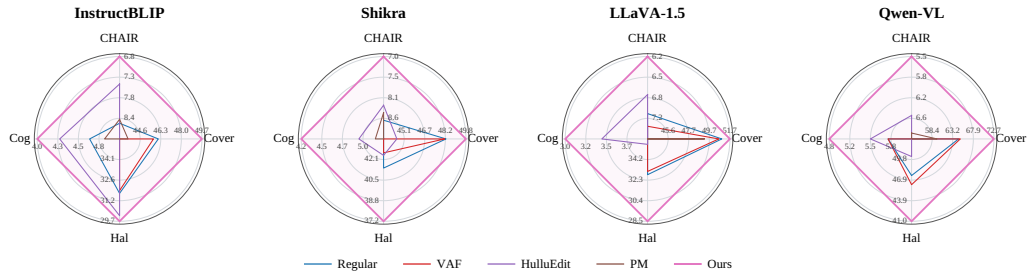


Figure 5: Comparison on AMBER using InstructBLIP, Shikra, LLaVA-1.5, and Qwen-VL.

E Ablation Study

Tables 4 and 5 report the ablation results of CIVER on CHAIR under maximum generation lengths of 64 and 128 tokens. Overall, the complete CIVER consistently achieves the best or tied-best performance across the four LVLMs, confirming that the proposed components contribute complementary information to hallucination mitigation. Compared with Regular decoding, the full model reduces C_S and C_I by an average of 8.65 and 3.13 points, respectively, under the 64-token setting. The reductions further increase to 14.55 and 5.30 points when the maximum generation length is extended to 128 tokens, indicating that the advantage of CIVER becomes more evident during longer generation.

Table 4: Ablation study of CIVER on CHAIR with a maximum generation length of 64 tokens. Lower C_S and C_I indicate fewer hallucinations.

	InstructBLIP		Shikra		LLaVA-1.5		Qwen-VL	
	C_S	C_I	C_S	C_I	C_S	C_I	C_S	C_I
Regular	25.8	9.7	24.7	8.7	22.0	6.4	24.1	7.8
w/o Visual Prior	16.8	7.8	17.5	8.2	14.3	7.9	16.1	6.8
w/o Candidate Competition	20.5	6.7	20.3	7.1	17.4	5.5	19.4	7.1
w/o Evidence Memory	20.1	6.5	20.1	6.9	17.8	5.8	18.3	6.5
w/o Ambiguity Gate	18.0	5.7	18.4	6.1	16.2	5.1	16.2	5.7
Ours	**16.1**	**5.1**	**16.8**	**5.4**	**14.3**	**4.7**	**14.8**	**4.9**

Table 5: Ablation study of CIVER on CHAIR with a maximum generation length of 128 tokens. Lower C_S and C_I indicate fewer hallucinations.

	InstructBLIP		Shikra		LLaVA-1.5		Qwen-VL	
	C_S	C_I	C_S	C_I	C_S	C_I	C_S	C_I
Regular	55.6	17.6	54.3	15.7	49.6	13.8	51.8	16.7
w/o Visual Prior	48.0	14.9	46.8	13.7	43.8	12.0	44.1	13.8
w/o Candidate Competition	51.4	15.4	50.2	14.1	46.4	12.3	47.5	14.3
w/o Evidence Memory	48.0	14.5	46.8	13.4	43.8	11.7	44.1	13.4
w/o Ambiguity Gate	51.7	15.1	50.4	13.8	46.6	12.1	47.8	14.0
Ours	**40.2**	**11.5**	**39.0**	**11.2**	**37.8**	**9.7**	**36.1**	**10.2**

Among the individual components, candidate competition plays a particularly important role. Removing it increases C_S and C_I by an average of 3.90 and 1.58 points at 64 tokens, and the degradation enlarges to 10.60 and 3.38 points at 128 tokens. This result supports the motivation of CIVER that simply identifying visual evidence associated with the current candidate is insufficient; explicitly modeling the relative evidence that favors competing candidates provides useful information for distinguishing visually supported predictions from plausible but unsupported alternatives. Removing the evidence memory also consistently degrades performance, with larger drops under the 128-token setting, suggesting that evidence accumulated across decoding steps provides an increasingly useful complementary reference as generation proceeds.

The visual prior and ambiguity gate are also important for stabilizing the refinement process. Without the visual prior, C_I increases noticeably across the four backbones, indicating that an input-dependent reference helps complement candidate evidence estimated from the current decoding

state. The effect becomes substantially stronger during longer generation, where removing the prior increases C_S and C_I by 7.40 and 2.95 points on average. Similarly, disabling the ambiguity gate produces relatively moderate degradation at 64 tokens but substantially larger degradation at 128 tokens, increasing C_S and C_I by 10.85 and 3.10 points on average. This suggests that uniformly applying evidence-based attention adjustment is less effective than adapting the intervention strength according to the ambiguity of the current candidate pair, particularly when decoding trajectories become longer and more diverse.

Taken together, these results show that CIVER does not rely on a single dominant component. Candidate-level competition provides the main discriminative signal, while the sample-specific visual prior and cross-step memory supply complementary references, and the ambiguity gate controls when stronger intervention is needed. Their combination yields more stable hallucination mitigation across different LVLM architectures and becomes especially beneficial in longer-form generation.

F Hyperparameter Analysis

We analyze four representative hyperparameters of CIVER on the CHAIR benchmark using LLaVA-1.5-7B. When varying one hyperparameter, the others are kept fixed to the reference configuration. Figure 6 reports CHAIR_I, where lower values indicate fewer object-level hallucinations.

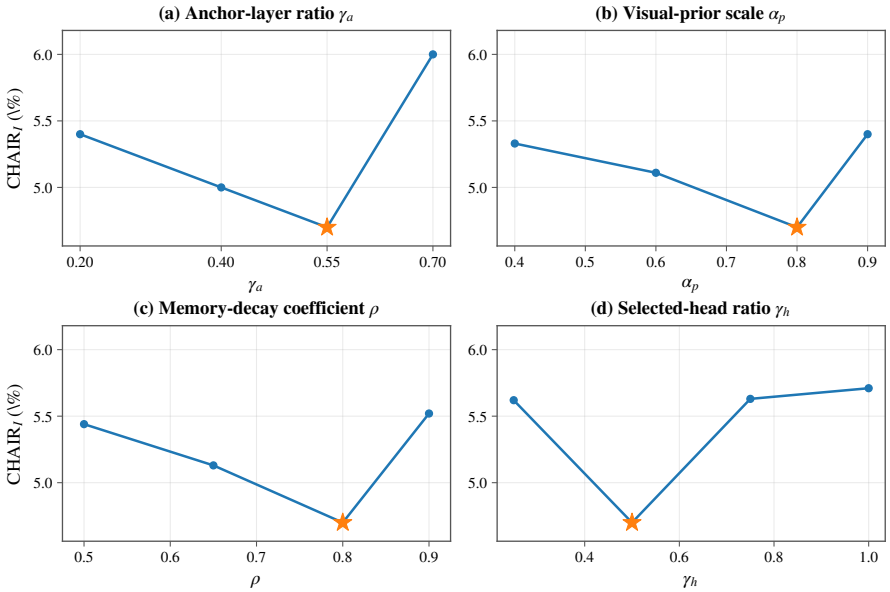


Figure 6: Sensitivity of CIVER to four representative hyperparameters on CHAIR using LLaVA-1.5-7B. Lower CHAIR_I indicates fewer object hallucinations.

Anchor-layer ratio.

As shown in Fig. 6(a), among the evaluated settings, CIVER achieves the lowest CHAIR_I of 4.70% at $\gamma_a = 0.55$. Shallower anchor-layer ratios of 0.20 and 0.40 yield CHAIR_I scores of 5.40% and 5.00%, respectively, while increasing the ratio to 0.70 raises the score to 6.00%. These results suggest that an intermediate anchor layer provides a favorable balance between sufficiently informative candidate representations and sufficient remaining depth for subsequent attention refinement.

Visual-prior scale.

Figure 6(b) shows that increasing α_p from 0.40 to 0.80 steadily improves hallucination mitigation, reducing CHAIR_I from 5.33% to 4.70%. Further increasing the scale to 0.90 degrades performance. These results indicate that the sample-specific visual prior provides a useful input-level reference, whereas an excessively strong prior may weaken the contribution of the dynamically estimated candidate evidence. We therefore set $\alpha_p = 0.80$.

Memory-decay coefficient.

As shown in Fig. 6(c), the best performance is obtained at $\rho = 0.80$, with $\mathrm{CHAIR}_I = 4.70\%$. Smaller values weaken cross-step continuity by placing more emphasis on the current evidence, whereas an excessively large value may overemphasize previously accumulated evidence. This result suggests that moderate memory persistence provides a useful balance between current and cross-step visual evidence.

Selected-head ratio.

Figure 6(d) shows that selecting half of the attention heads for adjustment achieves the best result, with $\mathrm{CHAIR}_I = 4.70\%$ at $\gamma_h = 0.50$. Selecting fewer heads limits the scope of evidence-guided adjustment, whereas increasing the ratio to 0.75 or 1.00 degrades performance. This result supports selectively adjusting evidence-aligned heads rather than indiscriminately modifying all attention heads.

Based on these sensitivity results, we set $\gamma_a = 0.55$, $\alpha_p = 0.80$, $\rho = 0.80$, and $\gamma_h = 0.50$ in all experiments. Overall, CIVER remains relatively stable within moderate parameter ranges, while excessively deep anchor placement, overly strong visual priors, excessive memory persistence, or overly broad head intervention can reduce its effectiveness.

G Candidate Accessibility and Oracle Analysis

Section 3 uses the native anchor-layer Top-2 pair as an inference-time probe of local candidate competition. This raises two related questions: how accessible visually grounded hypotheses are at the anchor layer, and whether replacing the native pair with a grounding-informed pair would improve the final prediction. We investigate these questions on the held-out CHAIR analysis split. Because open-ended captioning admits multiple valid continuations, we do not assume a unique gold next token. Instead, for each originally hallucinated object event, we use the CHAIR/COCO image annotations to define visually grounded object concepts and retain only contextual surface forms that can be realized as a single token under the model tokenizer. These annotation-derived candidates are used only for post-hoc diagnosis and the label-informed Grounded Oracle; they are not available to Native CIVER or any main comparison method.

Figure 7(a) first measures how much candidate rankings can still change after the anchor layer. Across the 1,704 eligible object events, the token eventually selected by the final layer appears in the anchor-layer Top-1, Top-2, Top-5, Top-10, and Top-50 sets with probabilities of 22.71%, 31.16%, 42.66%, 51.12%, and 68.60%, respectively. Candidate rankings at the intermediate layer therefore remain substantially different from the final decisions reached after the remaining decoder layers.

Grounded hypotheses are considerably less accessible at originally hallucinated positions. Among the 211 hallucinated events, the anchor candidate set contains at least one eligible grounded object candidate in only 3.79%, 4.74%, 8.06%, 9.95%, and 24.17% of the events at Top-1, Top-2, Top-5, Top-10, and Top-50, respectively. In particular, the low Top-2 coverage shows that CIVER's native pair should not be interpreted as a closed candidate set expected to contain a correct continuation. It instead provides a local probe of the competition state that informs subsequent evidence refinement while the final prediction remains unrestricted over the full vocabulary.

We next examine whether this accessibility is associated with the behavior of Native CIVER. Only 21 hallucinated events contain an eligible grounded candidate within the anchor-layer Top-10, so the Top-1, Top-2, ranks 3–5, and ranks 6–10 bins are merged according to a predefined minimum-group-size rule. As shown in Fig. 7(b), when a grounded candidate is accessible within Top-10, CIVER suppresses 33.33% of the original hallucinated tokens and changes 14.29% of the events to grounded object predictions. When the grounded candidate lies beyond Top-10 or no eligible contextual single-token realization is available, these rates decrease to 16.84% and 2.11%, respectively. Because the accessible group contains only 21 events and consequently has a wide confidence interval, this comparison should be interpreted as a descriptive effect pattern: CIVER appears more effective when a visually grounded hypothesis is already locally accessible.

To test whether native candidate selection is itself the dominant bottleneck, we construct a diagnostic Grounded Oracle-CIVER. It replaces the native Top-2 evidence pair with the highest-ranked eligible grounded object candidate and the strongest unsupported object candidate, while leaving all

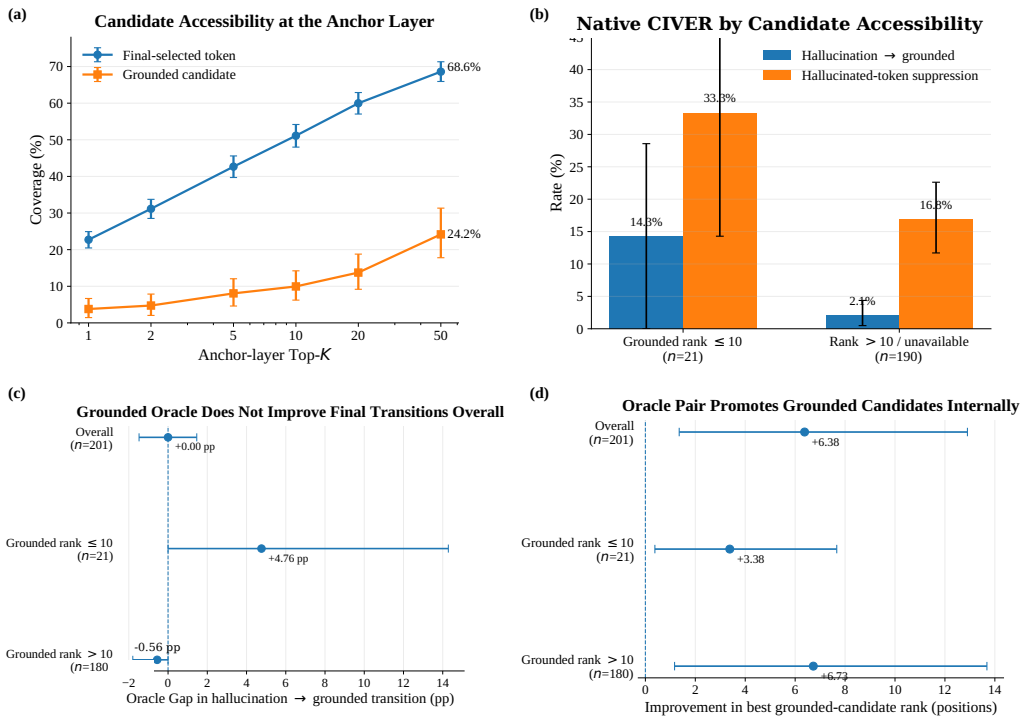


Figure 7: Candidate accessibility and Grounded Oracle analysis on CHAIR. (a) Anchor-layer Top-K coverage of the final selected token over 1,704 eligible object events and of at least one contextual single-token grounded candidate over 211 originally hallucinated events. The curves answer different questions and use different denominators. (b) Native CIVER outcomes when the highest-ranked grounded candidate has anchor rank ≤ 10 ($n = 21$) or has rank > 10 or is unavailable ($n = 190$). (c) Paired Grounded Oracle–Native difference in hallucination-to-grounded transition rate; positive values favor the diagnostic Oracle. (d) Improvement in the rank of the best grounded candidate, measured as Native rank minus Oracle rank. Panels (c)–(d) contain 201 paired events: 21 with grounded rank ≤ 10 and 180 with grounded rank > 10. The 10 events without an eligible grounded candidate are excluded because an Oracle pair cannot be constructed. Error bars denote 95% image-cluster bootstrap confidence intervals.

subsequent CIVER operations and the final full-vocabulary prediction unchanged. The 10 events without an eligible grounded candidate cannot form this pair and are excluded, leaving 201 paired events. Figure 7(c) reports the paired Oracle–Native difference in hallucination-to-grounded transitions. Across all paired events, the difference is 0.00 percentage points, with a 95% image-cluster bootstrap confidence interval of $[-1.48, 1.46]$. The point estimate is $+4.76$ points for the 21 events with grounded rank ≤ 10 (95% CI $[0, 14.29]$) and -0.56 points for the 180 events with grounded rank > 10 (95% CI $[-1.80, 0]$). Thus, replacing the native pair with a label-informed grounded pair does not yield a consistent improvement in the final grounded prediction.

The internal candidate ranking shows a different pattern. As shown in Fig. 7(d), Grounded Oracle-CIVER improves the rank of the best grounded candidate by 6.38 positions on average, with a 95% confidence interval of $[1.36, 12.90]$. The corresponding point estimates are $+3.38$ positions when the grounded candidate initially lies within Top-10 and $+6.73$ positions when it lies beyond Top-10; the latter has a 95% confidence interval of $[1.17, 13.68]$. A grounding-informed candidate pair can therefore promote a grounded hypothesis within the decoder without necessarily making it the final Top-1 prediction. This separation indicates that candidate accessibility is an important boundary condition, but candidate selection alone does not determine successful hallucination mitigation. Other aspects of evidence estimation and subsequent decoder evolution must also contribute to whether an internally promoted grounded hypothesis becomes the final output.

H Inference Efficiency Analysis

We further evaluate the inference efficiency of different decoding methods on CHAIR with the maximum generation length set to 64 tokens. Figure 8 reports the average decoding time per sample. Regular decoding requires 1.900 s, while CIVER takes 2.446 s, corresponding to an additional overhead of only 0.546 s per sample. Its runtime is also close to VAF (2.334 s), indicating that the candidate-informed evidence refinement introduces only moderate computational cost. In contrast, HulluEdit and PM require 7.440 s and 11.750 s per sample, respectively.

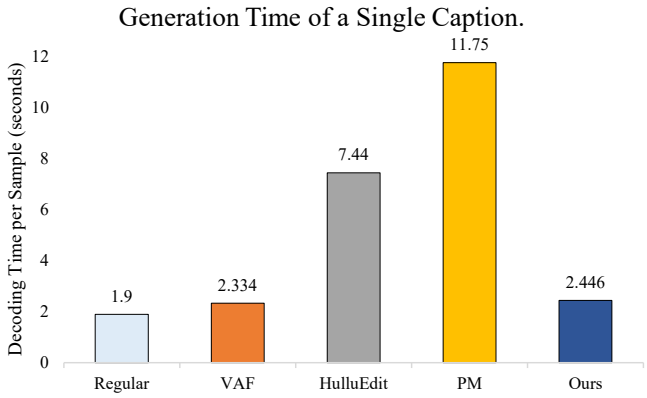


Figure 8: Average decoding time per sample on CHAIR with the maximum generation length set to 64 tokens. Lower values indicate higher inference efficiency.

Notably, CIVER is approximately $3.0\times$ faster than HulluEdit and $4.8\times$ faster than PM. These results show that CIVER achieves a favorable trade-off between hallucination mitigation and inference efficiency: although candidate comparison and evidence-guided attention adjustment introduce a small overhead over standard decoding, they avoid the substantially higher computational cost of competing decoding methods.

I Case Study

Fig. 9 presents four qualitative examples comparing CIVER with VAF, PM, and HulluEdit, where hallucinated content and visually grounded content are highlighted in red and blue, respectively. Across different scenes, the baseline methods frequently introduce plausible but visually unsupported objects. For example, they hallucinate a "clock," "book," "toy car," and "handbag" in the salon scene; additional cars, trucks, or people in the motorcycle-racing scene; unsupported vehicles and people around the train; and dogs on the boat. In contrast, CIVER generally focuses on visually supported scene elements and avoids these red-highlighted object hallucinations.

These examples cover diverse indoor, dynamic, urban, and water scenes, yet exhibit a similar pattern: hallucinated objects are often semantically compatible with the surrounding context but insufficiently supported by the image. CIVER is less likely to introduce such plausible but unsupported content, which is consistent with its candidate-informed design that explicitly compares the relative visual evidence of competing candidates during decoding.



Prompt: Please describe this image in detail.

Vaf: The image features a man cutting a child's hair in a salon. The child is sitting in a chair, and the man is skillfully using a pair of scissors to trim the child's hair. The man is focused on his task, ensuring the child's haircut is done well. In the background, there is a couch and a chair, providing a comfortable environment for the child and the man. **A clock** is also visible on the wall, allowing the man to keep track of time during the haircut.

PM: The image showcases a little boy sitting in a salon chair with a man cutting his hair. The boy appears to be excited and enjoying the process. The Barber is skillfully trimming the boy's hair using a scissors placed close to the child's head. The salon has chairs arranged all around the space; one is occupied by the child, another one is visible in the background, and multiple smaller chairs are scattered in the background. The room is also equipped with various recreation and leisure items, **including a book, a toy car.**

HulluEdit: The image depicts a man cutting a child's hair in a salon setting. The child is seated in a chair, and the man is skillfully using a pair of scissors to trim the child's hair. Another person is present in the scene, possibly assisting or observing the haircut.The salon is furnished with a couch and a couple of chairs, one of which the child is seated on. There is also a **handbag** placed on the floor near the couch. The salon appears to be well-equipped for haircuts and other...

Ours: The image features a man and a young child sitting in a chair, with the man cutting the child's hair. There are some posters on the wall behind the man. The scene appears to be taking place in a salon or a similar setting. In the background, there is a couch and a chair, providing a comfortable environment for the child and the man. **A woman is also present in the scene, possibly assisting the man or observing the haircut.**



Prompt: Please describe this image in detail.

Vaf:The image captures a thrilling moment on a race track, with two motorcycles racing side by side. The motorcycles are positioned close to each other, with one slightly ahead of the other. The riders are skillfully navigating the track, showcasing their expertise in motorcycle racing.In addition to the motorcycles, there are a few other objects in the scene. **A car** can be seen in the background, and **a truck** is parked further away. There are also two people present in the scene, likely spectators or staff members, observing the race.

PM: The image depicts a thrilling race on a race track with two motorcycles. The bikes are racing head-to-head, with their riders leaning low, carving up the track. One motorcycle is slightly ahead of the other, indicating the intense competition between the racers. **The track** features several barriers located alongside the racing strip. Some of the trucks are parked along the edge, likely awaiting to load equipment or support staff. Overall, the scene captures the exhilarating atmosphere of a professional motorsport event.

HulluEdit: The image captures a thrilling moment during a motorcycle race on a track. Two motorcycles are speeding down the track, with one motorcycle slightly ahead of the other. Both riders are wearing helmets for safety.In addition to the motorcycles, there are a few other objects in the scene. There is **a car** parked on the side of the track, and **a truck** can be seen in the background. **Two people** are also present in the scene, possibly spectators or race officials.",

Ours: The image captures a thrilling moment on a race track, with two motorcycles racing side by side. The motorcycles are positioned close to each other, with one slightly ahead of the other. The riders are wearing helmets and are focused on the race, giving the scene a sense of excitement and competition. The overall atmosphere of the image is one of speed and adrenaline, as the two motorcycles race down the track.",



Prompt: Please describe this image in detail.

VAF: The image features a red and white Virgin train traveling down the tracks, passing by a brick building. The train is positioned in the middle of the scene, with the brick building on its left side. In the background, there are **several cars and a bus visible**, likely indicating a busy urban area. **The cars are scattered around the scene**, with some closer to the train and others further away. The bus is located on the right side of the image, near the edge of the frame.

PM:The scene features a large red railroad vehicle, possibly a Virgin train, traveling through a town. The train is the main focus of the picture and is followed by **several cars on the road**. The train spans almost the entire length of the image. There are **multiple people** in the scene, with some of them walking on the side of the train and others standing nearby or traveling in **other cars** in the area. **Some cars** are closer to the train while others are situated further back. The atmosphere appears to be bustling and busy in this urban setting.

HulluEdit: The image depicts a red, white, and yellow train traveling down the tracks in a city setting. The train is positioned in the middle of the scene, surrounded by tall buildings on both sides of the tracks. In addition to the train, there are several cars and trucks scattered throughout the scene. **Some cars** can be seen on the left and right sides of the train, while **a few trucks** are parked or driving near the train tracks. The presence of **these vehicles** suggests that the train is passing through a busy urban area.

Ours: The image features a red and white Virgin train traveling down the tracks, passing by a brick building. The train is positioned in the middle of the scene, with the red and white colors of the train contrasting nicely with the surrounding environment. In the background, there are several cars and buses parked or driving, indicating that the train is passing through a busy urban area. The cars are scattered around the scene, with some closer to the train and others further away. A bus can be seen on the left side of the image, and another bus is located on the right side.



Prompt: Please describe this image in detail.

VAF: The image features a large orange boat floating on a body of water, possibly the ocean. There are two people on the boat, one standing closer to the front and the other near the middle. Both individuals are wearing blue shirts, and they appear to be enjoying their time on the boat. The boat is equipped with a ladder, which is located near the middle of the boat. The scene also includes a backpack, possibly belonging to one of the individuals on the boat. The overall atmosphere suggests a leisurely day out on the water.

PM: The image features a men's boating adventure with a pair of men operating a large orange watercraft, likely a speedboat or an inflatable tug boat, that is attached to a cable. They are engaged in towing and steering the boat across the water. In addition to the two men controlling the boat, there is also **a dog** on board, enjoying the ride with the men. Two more people can be seen in the background, but they are not actively participating in the boat's operation at this time.The orange watercraft covers a significant portion of the image.

HulluEdit: The image features a large orange boat floating on a body of water, possibly the ocean. There are several people on the boat, enjoying their time on the water. One person is standing near the front of the boat, while others are scattered throughout the vessel. In addition to the people, **there are two dogs on the boat as well. One dog is located near the middle of the boat, while the other dog is situated closer to the front.** The boat appears to be a ferry, providing transportation for the passengers and their pets.

Ours: The image features a large orange boat floating on the water, with two people standing on the deck. One person is closer to the left side of the boat, while the other is positioned more towards the center. They appear to be enjoying their time on the boat, possibly preparing for a rescue or simply enjoying the water. The boat is equipped with a winch, which is likely used for lifting or lowering equipment or personnel. The scene captures the essence of a boat ride, with the two people on board and the boat gliding through the water.

Figure 9: Qualitative comparison of CIVER with VAF, PM, and HulluEdit on four generation examples using LLaVA-1.5. Hallucinated content and visually grounded content are highlighted in red and blue, respectively.