ADAPTIVE VISUAL EVIDENCE-GUIDED DECODING FOR HALLUCINATION REDUCTION IN LVLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large Vision-Language Models (LVLMs) are prone to hallucinations, producing outputs that are linguistically plausible but not grounded in visual evidence. Existing mitigation approaches typically rely on multi-pass interventions, attention reallocation, or look-twice visual re-injection mechanisms. Although effective, these methods either incur significant computational overhead or provide only partial mitigation, limiting their reliability in practical deployment. In this work, we introduce **A**daptive **V**isual **E**vidence-Guided **D**ecoding (AVED), a unified single-forward-pass framework that explicitly incorporates visual evidence into the decoding process. Instead of treating visual grounding as an external correction step, AVED integrates visual evidence estimation and decoding within a single forward pass, eliminating the need for multiple inference stages or external visual re-injection. Specifically, our method consists of two complementary components: Visual Attention Calibration (VAC), which refines cross-modal attention to capture vision-language correspondence, and Layer-wise Logits Reweighting (LLR), which adaptively aggregates intermediate decoding predictions according to their estimated visual evidence reliability. The two components are tightly connected through cross-attention signals, where calibrated attention provides visual evidence estimates that guide evidence-aware logits reweighting. Extensive experiments on hallucination benchmarks demonstrate that our approach consistently suppresses hallucinations while improving generalization under distribution shifts, establishing a simple yet effective paradigm for reliable LVLM inference.

1 INTRODUCTION

Large Vision-Language Models (LVLMs) have recently demonstrated remarkable capabilities in multimodal understanding and generation across images and text [1–7]. Despite these advances, a persistent and critical limitation is *hallucination*, where models generate content that is not supported by the visual input but is instead inferred from language priors or dataset biases [8–12]. Such failures significantly hinder the deployment of LVLMs in safety-critical or high-reliability applications. Recent efforts to mitigate hallucinations typically rely on multi-pass deconfounding strategies [13–15], attention reallocation [16–21], and look-twice visual re-injection mechanisms [22–24]. Deconfounding approaches mainly include backdoor adjustment [13, 25, 26] and counterfactual reasoning [14, 15, 27]. While these methods provide a principled perspective on bias mitigation, they often require multiple forward passes, which limits their efficiency in practical deployment scenarios.

Although several single-forward-pass methods have been explored, they typically achieve only partial mitigation or may introduce additional noisy or misaligned signals. Attention reallocation methods dynamically adjust attention maps to improve cross-modal relevance. However, they often overlook biases that bypass the attention mechanism, which can arise from residual connections or deeper multimodal fusion pathways. Another line of work focuses on visual information re-injection [23, 24, 28], which reinforces visual signals during decoding. While effective within a single inference pass, these methods introduce additional visual injection pathways that may propagate noisy or inconsistent signals. In particular, since the selection of injected visual tokens depends on the evolving hidden states [23], any degradation in visual faithfulness within these hidden states can lead to suboptimal or noisy visual re-injection.

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

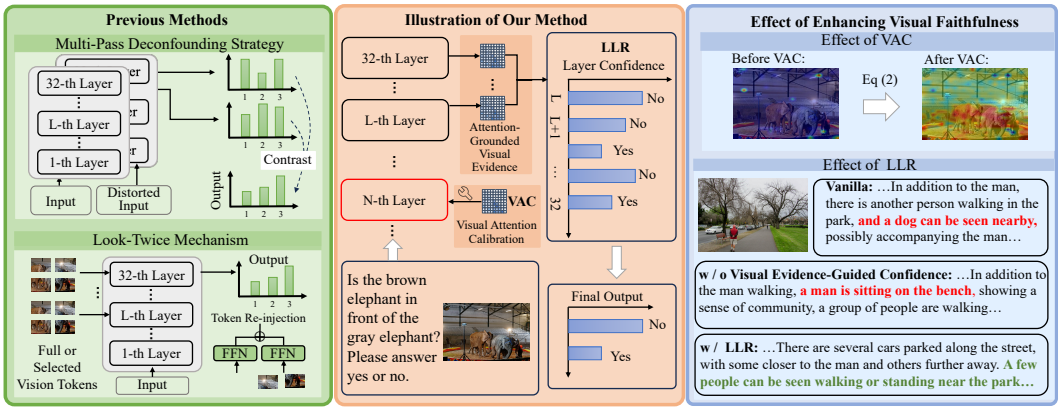

Figure 1: Comparison between prior methods and our approach. Existing hallucination mitigation strategies typically rely on multiple forward passes, attention reallocation, or look-twice visual re-injection, which can incur additional computational overhead or introduce misaligned signals. In contrast, our framework performs unified Adaptive Visual Evidence-Guided Decoding through two complementary components: Visual Attention Calibration (VAC) and Layer-wise Logits Reweighting (LLR). VAC supplies the attention-derived relative evidence signal used by LLR to compare intermediate layers.

In this work, we argue that hallucinations in LVLMs can be addressed from the perspective of Adaptive Visual Evidence-Guided Decoding, where visual grounding is not treated as an external correction step but is explicitly integrated into the decoding process. From this perspective, we propose a unified Adaptive Visual Evidence-Guided Decoding (AVED) framework for hallucination mitigation in LVLMs. AVED integrates visual evidence estimation directly into the decoding process within a single forward pass, eliminating the need for multi-pass inference or external visual re-injection. Instead of introducing additional visual signals, it leverages intrinsic cross-modal interactions to guide decoding with visual evidence in a unified manner.

Specifically, our method introduces two complementary mechanisms. First, Visual Attention Calibration (VAC) refines cross-modal attention distributions to align with vision-language correspondence, improving visual grounding at the attention level. Second, Layer-wise Logits Reweighting (LLR) adaptively aggregates intermediate layer predictions based on attention-grounded visual evidence consistency, enabling decoding decisions to emphasize visually faithful representations while suppressing spurious linguistic priors. Importantly, VAC and LLR are tightly connected through cross-attention-derived visual evidence signals, forming an adaptive evidence-guided decoding process.

AVED uses normalized visual similarity to allocate the preserved visual-attention mass. LLR then computes a relative concentration score over attention-selected visual tokens. These choices make the evidence signal selective and task-aware; the ablations in Section 4.3 evaluate the main alternatives.

AVED uses one autoregressive model evaluation per generated token and does not require a second model evaluation or externally injected visual tokens. It does require additional similarity and selected-row value computations within that evaluation, which is reflected in the runtime measurement. Our contributions are summarized as follows:

- We propose AVED, a unified Adaptive Visual Evidence-Guided Decoding framework that mitigates hallucinations in LVLMs without multi-pass inference or external visual re-injection. In addition to attention calibration, AVED adjusts the contribution of intermediate-layer predictions at the decoding stage.
- AVED establishes visual-evidence coupling between attention calibration and layer-wise decoding, together with a layer-confidence criterion grounded in visual evidence rather than linguistic confidence alone. Controlled comparisons with five alternative layer confidence measures substantiate the proposed layer-wise confidence criterion.
- Experiments show improvements over the compared methods on several reported benchmarks and perturbation levels.

2 Related Work

Numerous approaches have been proposed to mitigate hallucinations in LVLMs, which can be broadly categorized into attention reallocation [16–19], visual signal enhancement [28–30], causal reasoning [26, 27, 31–37], and decoding strategies [38–40]. Among these, causal reasoning methods, such as backdoor adjustment and counterfactual reasoning, and decoding-based strategies provide a principled framework for bias mitigation. However, they typically require multiple forward passes, limiting their efficiency in practical deployment. For instance, causal intervention-based decoding [41] explicitly adjusts the decoding process to suppress hallucinations, while counterfactual reasoning [14, 31, 42, 43] analyzes prediction shifts under hypothetical interventions to identify hallucination-prone tokens or modalities.

To improve efficiency, a growing body of work focuses on single-forward-pass methods. Despite their practicality, these approaches often provide only partial mitigation or introduce misaligned signals. For example, prior works [17, 18, 33, 44] attempt to enforce balanced cross-modal attention, aiming to prevent over-reliance on linguistic priors. Huang et al. [17] further observe that LVLMs tend to depend heavily on summary tokens, and propose a retrospection-based beam search strategy to penalize such dependencies. While effective to some extent, these methods primarily operate within the attention space and overlook biases that bypass attention mechanisms, such as those introduced through residual connections or multimodal fusion pathways.

Another line of work adopts the look-twice strategy to alleviate the degradation of visual information in deep decoding layers. Zou et al. [28] treat visual tokens as supplementary evidence and re-inject them into the feed-forward network, improving factual grounding. However, this visual enhancement method often injects tokens indiscriminately, introducing interference from background regions and distracting the model from salient cues. To address this issue, Zhu et al. [22] propose an Adaptive Visual Reinforcement (AIR) framework, which selectively integrates informative patches via prototype-based token reduction and optimal transport–guided reinforcement. Nevertheless, AIR primarily targets the decoding stage and does not address upstream attention biases; as a result, degraded hidden states induced by attention-level hallucinations may lead to suboptimal or noisy re-injection. For more comprehensive discussions, we refer readers to recent surveys [45–47].

3 Method

Unlike prior methods that rely on multi-pass inference [13–15] or visual re-injection [22, 28], our Adaptive Visual Evidence-Guided Decoding (AVED) integrates attention-based visual evidence estimation and visual evidence-guided decoding within a single forward pass, without requiring external visual signals. In this section, we first present the motivation behind AVED, followed by a detailed description of its two components.

3.1 Motivation for Adaptive Visual Evidence-Guided Decoding

For clarity, we first define the key symbols and notations used in our model implementation:

Notation. X_V denotes visual input (image), which is encoded into a sequence of visual tokens for model processing. X_T denotes textual input, serving as the contextual guidance for the model's generation task. We denote image tokens as $\mathbf{I}$ and text tokens as $\mathbf{T}$. $\mathbf{A}$ denotes the full attention tensor, whereas $\mathbf{A}^{\text{vis}}$ denotes its text-query-to-visual-key slice; we use lower-case bold symbols for vectors obtained by aggregating these tensors. This distinction is made explicit below. Y denotes the model output, representing the generated textual token.

Our key insight is that hallucinations often stem from insufficient utilization of visual evidence during autoregressive decoding. In practice, this manifests as two correlated effects: 1) *Attention Bias*: cross-modal attention is dominated by linguistic priors rather than visual evidence; 2) *Decoding Bias*: deeper decoder layers amplify language-model bias even when earlier layers contain visually grounded signals.

To address this within a unified single-pass framework, AVED introduces two complementary mechanisms: *Visual Attention Calibration (VAC)*, which refines cross-modal attention to reduce attention bias and better estimate reliable visual evidence from vision-language correspondence, and *Layer-wise*

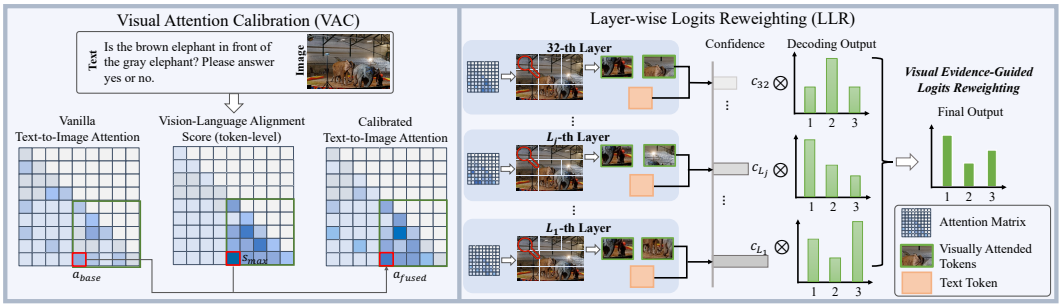

Figure 2: Implementation details of the proposed AVED method. Compared to the vanilla attention mechanism, Visual Attention Calibration (VAC) improves visual groundedness by refining cross-modal attention distributions to align with vision-language correspondence. Layer-wise Logits Reweighting (LLR) uses cross-modal attention to identify visually attended tokens and estimate a relative evidence concentration for each intermediate decoder layer. It then combines intermediate-layer logits according to these scores.

Logits Reweighting (LLR), which mitigates decoding bias by adaptively aggregating intermediate-layer predictions based on the attention-derived visual evidence. Together, these components enable an adaptive evidence-guided decoding process, where attention-based visual evidence directly guides layer-wise decoding aggregation, promoting visually grounded generation while suppressing reliance on spurious linguistic shortcuts.

3.2 Visual Attention Calibration

To reduce attention bias while preserving a valid attention distribution, Visual Attention Calibration (VAC) redistributes only the visual portion of selected attention rows. Let $|H|$, $|T|$, and $|I|$ denote the numbers of heads, text queries, and visual tokens, and let D be the hidden dimension. At a selected layer, the text-query-to-visual-key slice is $\mathbf{A}^{\text{vis}} \in \mathbb{R}^{|H|\times|T|\times|I|}$. VAC is applied at the layers and heads selected in the experimental setup[1]; all entries outside the selected rows remain unchanged. Let T and V be row-wise ℓ_2-normalized text and visual representations. We compute

$$S = TV^\top, \qquad s_j = \sum_{q=1}^{|T|} S_{q,j}, \qquad p_j = \frac{\exp(s_j/\tau)}{\sum_{i=1}^{|I|}\exp(s_i/\tau)},$$

where $\tau > 0$. Because p is a softmax of normalized similarities, it is nonnegative and sums to one. For each selected head r and query q, let the original visual mass be $m_{r,q} = \sum_{j=1}^{|I|} A^{\text{vis}}[r,q,j]$. VAC replaces the visual slice for that row with $A^{\text{cal}}[r,q,j] = m_{r,q}p_j$. We then mix the calibrated and original visual allocations:

$$A^{\text{fused}}[r,q,:] = \alpha A^{\text{cal}}[r,q,:] + (1-\alpha)A^{\text{vis}}[r,q,:], \qquad \alpha \in [0,1]. \tag{1}$$

Both terms have visual mass $m_{r,q}$, so the fused row remains nonnegative and preserves the original total attention mass after non-visual keys are left unchanged. VAC is therefore a mass-preserving redistribution of existing attention.

In the implementation, the similarity allocation is computed from hidden states already available at the current layer, and the value aggregation for the selected rows uses the fused weights before the next transformer block. Thus "single forward pass" means one autoregressive model evaluation per generated token. It does not mean that every internal attention operation is executed only once; the additional similarity and selected-row value computation is part of that same evaluation and is reflected in the runtime measurement.

3.3 Layer-wise Logits Reweighting

In this paper, we view intermediate decoding layers as hierarchical pathways, where different layers exhibit varying degrees of consistency with vision-language alignment. Based on this insight, we

[1]We follow prior work [48] to identify hallucination-prone layers.

Algorithm 1 Visual Attention Calibration

Require: Hidden states $\mathbf{h}$, post-softmax attention $\mathbf{A}$, image positions $\mathcal{I}$, text-query positions $\mathcal{T}$, selected heads $\mathcal{H}_a$, temperature τ, and mixing coefficient α.

Extract $V \leftarrow \mathbf{h}[\mathcal{I}, :]$ and $T \leftarrow \mathbf{h}[\mathcal{T}, :]$ and normalize each row.
Extract $A^{\text{vis}} \leftarrow \mathbf{A}[:, \mathcal{T}, \mathcal{I}]$.
Compute $S \leftarrow TV^\top$, $s_j \leftarrow \sum_q S_{q,j}$, and $p_j \leftarrow \text{softmax}(s_j/\tau)$.
for each selected head $r \in \mathcal{H}_a$ and query $q \in \mathcal{T}$ **do**
    $m_{r,q} \leftarrow \sum_j A^{\text{vis}}[r, q, j]$.
    $A^{\text{cal}}[r, q, j] \leftarrow m_{r,q} p_j$ for each visual key j.
    $A^{\text{vis}}[r, q, :] \leftarrow \alpha A^{\text{cal}}[r, q, :] + (1 - \alpha) A^{\text{vis}}[r, q, :]$.
end for
Recompute value aggregation for the modified rows and pass the result to the next block.
return Updated hidden states.

Algorithm 2 Layer-wise Logits Reweighting

Require: Visual features $\{\mathbf{v}_j\}_{j=1}^{|I|}$, selected layer set $\mathcal{L}$, current decoder hidden states $\{\mathbf{H}_l\}_{l\in\mathcal{L}}$, calibrated visual-attention slices $\{\mathbf{A}_l^{\text{vis}} \in \mathbb{R}^{|H|\times|T|\times|I|}\}_{l\in\mathcal{L}}$, LM head $f(\cdot)$, final-layer logits Y_L, top-K value K, temperature $\tau_v > 0$, and scaling factor β

Initialize confidence list $\mathcal{C} \leftarrow \emptyset$
Initialize logits list $\mathcal{Y} \leftarrow \emptyset$
for each layer $l \in \mathcal{L}$ **do**
    Aggregate calibrated attention over heads and text queries: $\bar{\mathbf{a}}_l \leftarrow \frac{1}{|H||T|}\sum_{r=1}^{|H|}\sum_{q=1}^{|T|}\mathbf{A}_l^{\text{vis}}[r, q, :] \in \mathbb{R}^{|I|}$
    Select top-K visually attended token indices: $\mathcal{I}_l \leftarrow \text{TopK}(\bar{\mathbf{a}}_l, K) \subseteq \{1, \ldots, |I|\}$
    Extract the task-aware decoding representation: $\mathbf{h}_l \leftarrow \mathbf{H}_l[-1]$ for captioning, and $\mathbf{h}_l \leftarrow |\mathcal{Q} \cup \mathcal{Y}_{\le t}|^{-1}\sum_{j\in\mathcal{Q}\cup\mathcal{Y}_{\le t}}\mathbf{H}_l[j]$ for VQA
    Compute normalized alignment score: $s_{l,j} = \frac{\mathbf{h}_l^\top \mathbf{v}_j}{\|\mathbf{h}_l\|_2\|\mathbf{v}_j\|_2}, \quad j \in \mathcal{I}_l$
    Compute layer confidence: $c_l = \max\left(\text{Softmax}(\{s_{l,j}/\tau_v\}_{j\in\mathcal{I}_l})\right)$
    Append confidence score: $\mathcal{C} \leftarrow \mathcal{C} \cup \{c_l\}$
    Compute intermediate logits: $Y_l = f(\mathbf{H}_l[-1])$
    Append logits: $\mathcal{Y} \leftarrow \mathcal{Y} \cup \{Y_l\}$
end for
Compute normalized layer weights: $w_l = \frac{\exp(c_l)}{\sum_{k\in\mathcal{L}}\exp(c_k)}$
Apply logits reweighting: $\hat{Y} = Y_L + \beta \cdot \sum_l w_l \cdot Y_l$
Return $\hat{Y}$

propose to use attention-derived visual evidence to estimate layer-wise confidence, such that VAC-calibrated evidence can improve the visual faithfulness of adaptive decoding. Let $\mathbf{h}_l$ denote the hidden representation at the l-th decoding layer, where $l \in \mathcal{L}$ and $\mathcal{L}$ is the set of layers considered for aggregation. We define a normalized layer-confidence distribution over $\mathcal{L}$ as

$$q_l \equiv q(l \mid X_V, X_T) = \frac{\exp(c_l)}{\sum_{k\in\mathcal{L}}\exp(c_k)}, \tag{2}$$

where c_l measures the consistency of layer l with attention-derived visual evidence. Specifically, for each layer l, let $\mathbf{A}_l^{\text{vis}} \in \mathbb{R}^{|H|\times|T|\times|I|}$ denote the calibrated text-query-to-visual-key attention slice produced by VAC. We aggregate it over heads and text queries to obtain one visual-token score vector,

$$\bar{\mathbf{a}}_l = \frac{1}{|H||T|}\sum_{r=1}^{|H|}\sum_{q=1}^{|T|}\mathbf{A}_l^{\text{vis}}[r, q, :] \in \mathbb{R}^{|I|}.$$

We then extract a subset of *visually attended tokens* Z_l^{attend} by selecting the top-K visual tokens according to $\bar{\mathbf{a}}_l$.

Why attention-guided top-K tokens? Restricting confidence estimation to the most attended visual tokens focuses the score on evidence relevant to the current input. Using all visual tokens can dilute this signal with background or weakly relevant content. Table 19 empirically supports our design.

Task-aware layer representation. For image captioning, we use the current decoding state $\mathbf{h}_l^{\text{cap}} = \mathbf{H}_l[-1]$. For VQA at decoding step t, we instead average only the question tokens $\mathcal{Q}$ and the generated answer tokens $\mathcal{Y}_{\le t}$:

$$\mathbf{h}_l^{\text{VQA}} = \frac{1}{|\mathcal{Q} \cup \mathcal{Y}_{\le t}|} \sum_{j \in \mathcal{Q} \cup \mathcal{Y}_{\le t}} \mathbf{H}_l[j]. \tag{3}$$

This question–answer mean captures both the visual information requested by the question and the evolving semantics of the answer. It avoids the recency sensitivity of a last-token state and the task-irrelevant system, formatting, or contextual tokens included by an all-text mean; Table 18 validates this choice.

We then compute normalized similarity between the layer representation $\mathbf{h}_l$ and each selected visual token $v_j \in Z_l^{\text{attend}}$:

$$\bar{s}_{l,j} = \frac{\mathbf{h}_l^\top v_j}{\|\mathbf{h}_l\|_2 \|v_j\|_2}. \tag{4}$$

The layer-wise visual-evidence score is defined as the largest normalized similarity probability over the attended visual tokens:

$$c_l = \max\Big(\text{Softmax}(\{\bar{s}_{l,j}/\tau_v\}_{v_j \in Z_l^{\text{attend}}})\Big), \tag{5}$$

where $\tau_v > 0$. VEC is a relative concentration signal: it measures how concentrated the normalized layer-to-visual similarity is among the attention-selected tokens. It is used to compare layers on the same input; it is not a calibrated probability of visual correctness and does not by itself certify object identity.

Why visual evidence-guided confidence? We refer to c_l as visual evidence-guided confidence (VEC). Unlike maximum logit, maximum probability, entropy, or logit-margin scores, VEC measures whether a layer representation is consistent with the visual tokens emphasized by cross-modal attention. This provides an attention-conditioned visual similarity signal that complements confidence measures derived from language logits; it does not establish whether a prediction is visually correct. The controlled comparison in Table 6 shows that this grounding-aware signal achieves the lowest sentence- and instance-level hallucination scores among the compared measures in this setting when coupled with VAC.

LLR operates in logit space. Let L denote the final decoder layer, $Y_L = f(\mathbf{H}_L[-1])$ its logits, and $Y_l = f(\mathbf{H}_l[-1])$ the logits of layer $l \in \mathcal{L}$. Define

$$w_l = \frac{\exp(c_l)}{\sum_{k \in \mathcal{L}} \exp(c_k)}. \tag{6}$$

The final logits and predictive distribution are

$$\hat{Y} = Y_L + \beta \sum_{l \in \mathcal{L}} w_l Y_l, \qquad p(Y \mid X_V, X_T) = \text{softmax}(\hat{Y}). \tag{7}$$

Thus LLR adds a weighted intermediate-layer logit ensemble to the final-layer logits before applying one softmax; it is not a mixture of per-layer probability distributions. The evidence for this aggregation is empirical and is evaluated by the ablations and benchmark results below.

We detail the full procedure in **Algorithm 2**. LLR uses attention-derived relative evidence to adapt the contribution of intermediate layers within the same autoregressive model evaluation.

4 EXPERIMENTS

In this section, we present the experimental results of our method across hallucination and general benchmarks, including performance comparisons, ablation studies, and visualization analyses.

4.1 Experiment Setup

Base Models and Baselines Alongside the proposed AVED, we evaluate several baseline approaches from prior literature, including vanilla LVLMs with standard greedy decoding, as well as state-of-the-art approaches such as OPERA [17], VCD [49], DoLA [38], LURE [50], HALC [39], ICD [51], MemVR [28], and AIR [22]. In particular, we compare AVED with ICD [51] and VCD [49], which adopt multi-pass contrastive decoding for hallucination mitigation. We further compare AVED with MemVR [28] and AIR [22], which follow a look-twice strategy over visual tokens to enhance factual grounding. All baseline methods are evaluated using the default configurations reported in the original papers. We conduct experiments on widely used LVLM backbones, namely LLaVA-1.5-7B [52] and Qwen2.5-VL-7B-Instruct [2]. We further evaluate the generalization ability of our method on Qwen3.5-4B [7] and GLM-4V-9B [53], with results shown in Tables 2 and 22. Baseline configurations and AVED hyperparameters are described in Appendix B.2.

Datasets and Metrics We evaluate our method on a variety of benchmarks. For hallucination assessment, we use CHAIR [54] on the MSCOCO [55] validation set and POPE (Polling-based Object Probing Evaluation) [11] on both clean and perturbed MSCOCO [55]. For general-purpose evaluation, we assess AVED on a range of visual question answering benchmarks, including MM-Vet [56], MME [57], and MMBench [58]. For evaluation metrics, we report the CHAIR metric [54], which is decomposed into sentence-level hallucination (CHAIR_S) and instance-level hallucination (CHAIR_I). We report accuracy and F1-score for POPE, task scores for MME, evaluator scores for MM-Vet, and accuracy for MMBench. Additional details on evaluation metrics are provided in Appendix B.1.

Table 1: Quantitative evaluation on the CHAIR benchmark. CHAIR_S and CHAIR_I (lower is better) measure sentence-level and instance-level hallucination severity, respectively. Recall (higher is better) measures the recall of correct objects. All methods are evaluated with a maximum of 1024 generated tokens.

Method	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	Recall ↑
LLaVA-1.5-7B	52.80 ↑0.0	13.94 ↑0.0	77.31 ↑0.0
DoLA	54.40 ↑1.60	14.33 ↑0.39	77.61 ↑0.30
OPERA	52.00 ↓0.80	14.90 ↑0.96	76.95 ↓0.36
VCD	57.60 ↑4.80	16.33 ↑2.39	78.03 ↑0.72
LURE	48.00 ↓4.80	13.78 ↓0.16	70.35 ↓6.96
HALC	51.03 ↓1.77	13.22 ↓0.72	72.28 ↓5.03
MemVR	47.40 ↓5.40	13.18 ↓0.76	**80.00** ↑2.69
Ours	**44.40** ↓8.40	**12.19** ↓1.75	<u>79.41</u> ↑2.10

4.2 Experiment Results

Results on CHAIR Benchmark We evaluate performance on the CHAIR benchmark based on LLaVA-1.5, with results reported in Table 1. Compared with the vanilla LLaVA-1.5 model, AVED reduces CHAIR_S and CHAIR_I by 8.4 and 1.75 percentage points, respectively, while Recall increases by 2.1 points. It obtains the lowest CHAIR values among the methods in this table.

Results on Clean and Perturbed POPE We report results on the clean POPE benchmark using LLaVA-1.5 and Qwen2.5-VL, with full results provided in Tables 7 and 8 in the Appendix. AVED improves the reported POPE scores on both evaluated backbones; these results support transferability across the tested models but do not establish backbone-independent behavior. In addition to results on the clean POPE benchmark, we further evaluate the robustness of AVED under distribution shifts. Following prior work [59], we construct a perturbed POPE setting by injecting Gaussian noise into image inputs. This controlled corruption tests sensitivity to additive image noise. Specifically, we first normalize pixel values and then add Gaussian noise with standard deviations of 0.08, 0.12, 0.18, 0.26, and 0.38, corresponding to progressively increasing perturbation severity.

Table 2: Comparison with the look-twice visual re-injection methods (such as MemVR and AIR) on MME-Hall, MME-Overall, and MMBench.

	Methods	MME-Hall ↑	MME-Overall ↑	MMBench ↑
LLaVA-1.5	Vanilla	643.3 ↑0.0	1863.9 ↑0.0	64.6 ↑0.0
	VCD	613.3 ↓30.0	1822.0 ↓41.9	61.6 ↓3.0
	VAF	603.3 ↓40.0	1746.6 ↓117.3	14.8 ↓49.8
	MemVR	648.3 ↑5.0	1891.0 ↑27.1	64.6 ↑0.0
	AIR	638.3 ↓5.0	1876.7 ↑12.8	64.7 ↑0.1
	Ours	**660.0** ↑16.7	**1900.4** ↑36.5	**65.1** ↑0.5
GLM-4V-9B	Vanilla	703.3 ↑0.0	2161.3 ↑0.0	81.3 ↑0.0
	VCD	696.7 ↓6.6	2153.7 ↓7.6	80.2 ↓1.1
	VAF	703.3 ↑0.0	2161.5 ↑0.2	81.3 ↑0.0
	MemVR	703.3 ↑0.0	2161.3 ↑0.0	81.3 ↑0.0
	AIR	703.3 ↑0.0	2161.5 ↑0.2	81.3 ↑0.0
	Ours	**713.3** ↑10.0	**2196.3** ↑35.0	**83.4** ↑2.1

Table 3: Average POPE performance under different perturbation levels based on Qwen2.5-VL. We report Accuracy and F1-score averaged over Random, Popular, and Adversarial settings.

Perturbation Levels	0.00		0.08		0.12		0.18		0.26		0.38	
	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑	Acc ↑	F1 ↑
Qwen2.5-VL-7B	87.69	86.58	85.92	84.37	84.68	82.76	82.37	79.75	80.07	76.74	75.89	71.42
OPERA	87.80	86.83	85.99	84.46	85.21	83.54	82.55	80.11	80.47	76.05	75.99	71.71
ICD	88.80	88.11	87.74	87.09	85.83	85.20	84.78	84.24	82.07	82.01	77.45	75.80
VCD	87.80	86.78	85.93	84.50	84.66	82.89	82.55	80.19	80.32	77.51	76.29	72.92
MemVR	88.01	86.97	86.14	84.67	85.06	83.31	83.53	81.39	80.55	77.71	76.72	72.77
Ours	**89.39**	**88.84**	**87.96**	**87.16**	**86.92**	**86.72**	**85.36**	**84.31**	**83.18**	**82.12**	**79.83**	**79.42**

Table 4: Comprehensive evaluation of recent methods on general visual question answering based on LLaVA-1.5 and Qwen2.5-VL. We report absolute performance as well as absolute changes from the baseline.

Evaluation	Methods	MME-Hall	Object-Level		Attribute-Level		MME-Overall	MM-Vet	MMBench
		Total ↑	Existence ↑	Count ↑	Position ↑	Color ↑	Total ↑	Score ↑	Accuracy ↑
LLaVA-1.5	Baseline	643.3 ↑0.0	190.0 ↑0.0	150.0 ↑0.0	128.3 ↑0.0	170.0 ↑0.0	1864.7 ↑0.0	31.10 ↑0.0	62.80 ↑0.0
	OPERA	610.0 ↓33.3	**195.0** ↑5.0	128.3 ↓21.7	121.7 ↓6.6	165.0 ↓5.0	1784.3 ↓80.3	32.00 ↑0.9	62.80 ↑0.0
	ICD	583.3 ↓60.0	185.0 ↓5.0	130.0 ↓20.0	121.7 ↓6.6	146.7 ↓23.3	1594.8 ↓269.9	25.90 ↓5.2	39.78 ↓23.0
	VCD	648.3 ↑5.0	190.0 ↑0.0	155.0 ↑5.0	128.3 ↑0.0	175.0 ↑5.0	1872.9 ↑8.2	30.20 ↓0.9	51.50 ↓11.3
	MemVR	648.3 ↑5.0	190.0 ↑0.0	155.0 ↑5.0	133.3 ↑5.0	**170.0** ↑0.0	1896.7 ↑32.0	**32.40** ↑1.3	63.75 ↑1.0
	AVED (Ours)	**660.0** ↑16.7	190.0 ↑0.0	**160.0** ↑10.0	**140.0** ↑11.7	**170.0** ↑0.0	**1900.4** ↑35.7	32.30 ↑1.2	**65.10** ↑2.3
Qwen2.5-VL	Baseline	705.0 ↑0.0	195.0 ↑0.0	165.0 ↑0.0	155.0 ↑0.0	190.0 ↑0.0	2322.7 ↑0.0	68.6 ↑0.0	80.67 ↑0.0
	OPERA	690.0 ↓15.0	190.0 ↓5.0	160.0 ↓5.0	155.0 ↑0.0	185.0 ↓5.0	2262.5 ↓60.2	67.2 ↓1.4	80.24 ↓0.43
	VCD	690.0 ↓15.0	**200.0** ↑5.0	150.0 ↓15.0	155.0 ↑0.0	185.0 ↓5.0	2283.9 ↓38.8	67.9 ↓0.7	80.58 ↓0.09
	ICD	700.0 ↓5.0	190.0 ↑0.0	**165.0** ↑0.0	160.0 ↑5.0	185.0 ↓5.0	2255.4 ↓67.3	67.0 ↓1.6	79.04 ↓1.63
	MemVR	705.0 ↑0.0	195.0 ↑0.0	**165.0** ↑0.0	155.0 ↑0.0	**190.0** ↑0.0	2322.7 ↑0.0	69.0 ↑0.4	83.76 ↑3.09
	AVED (Ours)	**713.3** ↑8.3	**200.0** ↑5.0	160.0 ↓5.0	**163.3** ↑8.3	**190.0** ↑0.0	**2338.5** ↑15.8	**69.5** ↑0.9	**84.36** ↑3.69

Table 5: Ablation study of VAC, LLR, and the visual evidence-guided confidence (VEC) in Equation 5.

Method	$CHAIR_S$ ↓	$CHAIR_I$ ↓	Recall ↑
LLaVA-1.5-7B	52.80 ↑0.0	13.94 ↑0.0	77.31 ↑0.0
w / o LLR	49.20 ↓3.60	13.85 ↓0.09	77.58 ↑0.27
w / o VAC	51.40 ↓1.40	14.32 ↑0.38	78.60 ↑1.29
w / o softmax	47.80 ↓5.00	13.71 ↓0.23	76.91 ↓0.40
w / o VEC	49.60 ↓3.20	13.81 ↓0.13	75.89 ↓1.42
Ours	**44.40** ↓8.40	**12.19** ↓1.75	**79.41** ↑2.10

Table 6: Comparison of visual evidence-guided confidence (VEC) with alternative layer-confidence measures on CHAIR.

Confidence measure	$CHAIR_S$ w/o VAC	$CHAIR_I$ w/o VAC	$CHAIR_S$ w/ VAC	$CHAIR_I$ w/ VAC
VEC (Ours)	32.8	**8.6**	**26.0**	**7.4**
Maximum logit	30.4	12.2	30.0	9.6
Maximum probability	29.4	10.4	29.8	7.8
Entropy	39.2	10.9	32.0	17.5
Uniform	42.8	16.1	40.4	10.8
Logit margin	**27.8**	19.3	27.8	13.7

As shown in Table 3, AVED achieves consistent improvements under all noise levels, with gains of up to 8.00 percentage points in average F1-score on the most severely perturbed setting (0.38). In particular, by leveraging intermediate-layer outputs, LLR adaptively reweights logits according to layer confidence, thereby steering predictions toward more visually grounded semantics and improving robustness under corrupted visual inputs. Detailed results for each noise level are provided in Tables 9-13 in Appendix B.3.

Results on General Visual Question Answering We evaluate the effectiveness of AVED on general visual question answering benchmarks, including MM-Vet, MME, and MMBench. The results are reported in Table 4. Overall, AVED outperforms or achieves comparable performance to competing methods across all evaluated benchmarks. Importantly, we further compare AVED with look-twice visual re-injection methods (e.g., MemVR and AIR) on MME and MMBench, with results presented in Table 2. It is observed that our method achieves the best performance on both LLaVA-1.5 and GLM-4V backbones, supporting its effectiveness on these benchmarks without requiring visual re-injection.

4.3 Ablation Studies and Visualization

Ablation Study Results We provide ablation study results for the proposed VAC, LLR, and visual evidence-guided confidence on the CHAIR benchmark in Table 5. VAC alone reduces both CHAIR metrics, whereas LLR alone reduces $CHAIR_S$ but increases $CHAIR_I$ relative to the baseline. Combining the two yields the lowest CHAIR scores among the variants in this table, suggesting complementary effects. Furthermore, we conduct additional ablations on the softmax operation in Equation 2 and the proposed visual evidence-guided confidence (denoted as VEC). The best

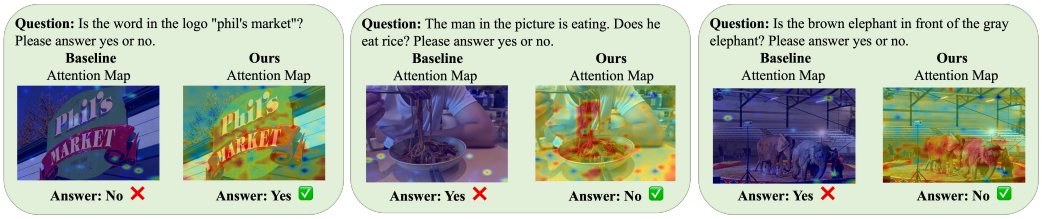

Figure 3: Effect of Visual Attention Calibration (VAC). AVED refines cross-modal attention by reallocating focus toward visually relevant regions, thereby improving the reliability of visual evidence estimation in LLR and reducing hallucinations during decoding.

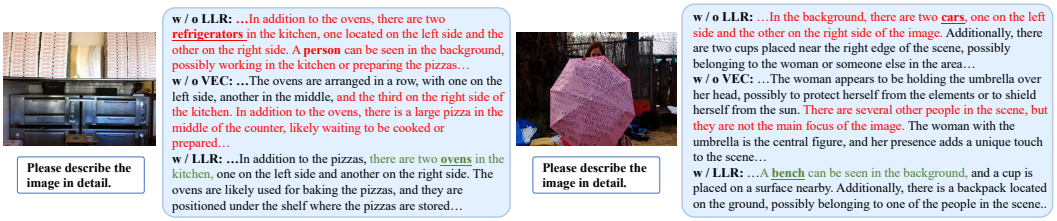

Figure 4: Effect of Layer-wise Logits Reweighting (LLR). We denote AVED without LLR as w/o LLR, and AVED without the visual evidence-guided confidence (i.e., using uniform layer confidence) as w/o VEC, while the full model is AVED. Hallucinated contents are highlighted in red.

CHAIR performance of our method highlights the effectiveness of incorporating visual evidence into confidence estimation for more accurate adaptive decoding.

Effect of Visual Evidence-Guided Confidence. Table 6 compares VEC with maximum logit, maximum probability, entropy, uniform weighting, and the top-1/top-2 logit margin. With VAC, VEC obtains the best $\text{CHAIR}_S/\text{CHAIR}_I$ pair, $26.0/7.4$. Some logit-derived measures perform well on one metric without VAC—for example, the logit margin reaches $\text{CHAIR}_S = 27.8$ but degrades CHAIR_I to 19.3—yet their gains are not consistent across both hallucination measures. Uniform weighting is also markedly worse, showing that the benefit is not merely an ensemble effect from averaging intermediate logits. These results support the utility of attention-conditioned visual similarity for layer weighting in the evaluated setting.

Effect of VAC We provide visualizations of both the vanilla attention and the calibrated attention to demonstrate that the proposed VAC module effectively grounds text-aligned visual information. The results are presented in Figure 3. It is shown that the baseline model exerts sparse visual attention. In contrast, AVED incorporates VAC to reallocate attention weights in a manner that better aligns with the text query, thereby enhancing visual faithfulness and mitigating attention degradation on visual tokens. The quantitative ablations and benchmark results provide the primary evidence for the effect of VAC; the visualization is qualitative.

Effect of LLR We provide case studies to illustrate the behavior of LLR using Equation 7. A comparison with the baseline method without LLR and visual evidence-guided confidence is provided in Figure 4. Additional case studies are presented in Figures 5-11. These examples show changes in intermediate-layer logits that are consistent with the reported benchmark improvements; they are qualitative illustrations rather than independent evidence of visual correctness.

5 Conclusion

This work advances the reliability of LVLMs by introducing a scalable and principled framework for hallucination mitigation. We propose Adaptive Visual Evidence-Guided Decoding (AVED), which integrates Visual Attention Calibration (VAC) to improve visual evidence estimation by refining cross-modal attention, together with Layer-wise Logits Reweighting (LLR), which adaptively aggregates intermediate-layer predictions based on attention-derived visual evidence, suppressing non-visual biases and improving decoding reliability. Extensive experiments on standard hallucination benchmarks demonstrate that AVED achieves superior hallucination suppression and robustness under distribution shifts.

REFERENCES

[1] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[2] 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*, 2024.

[3] 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*, 2023.

[4] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023.

[5] 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 *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198, 2024.

[6] Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, Jiazheng Xu, Bin Xu, Juanzi Li, Yuxiao Dong, Ming Ding, and Jie Tang. Cogvlm: Visual expert for pretrained language models. 2023.

[7] Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

[8] 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*, 2024.

[9] Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. *arXiv preprint arXiv:2411.11919*, 2024.

[10] Wei Lan, Wenyi Chen, Qingfeng Chen, Shirui Pan, Huiyu Zhou, and Yi Pan. A survey of hallucination in large visual language models. *arXiv preprint arXiv:2410.15359*, 2024.

[11] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023.

[12] Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063, 2024.

[13] Miao Yu, Zhenhong Zhou, Moayad Aloqaily, Kun Wang, Biwei Huang, Stephen Wang, Yueming Jin, and Qingsong Wen. Backdoor attribution: Elucidating and controlling backdoor in language models. *arXiv preprint arXiv:2509.21761*, 2025.

[14] Guanyu Zhou, Yibo Yan, Xin Zou, Kun Wang, Aiwei Liu, and Xuming Hu. Mitigating modality prior-induced hallucinations in multimodal large language models via deciphering attention causality. *arXiv preprint arXiv:2410.04780*, 2024.

[15] Junho Kim, Kim Yeonju, and Yong Man Ro. What if...?: Thinking counterfactual keywords helps to mitigate hallucination in large multi-modal models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 10672–10689, 2024.

[16] Yu Wang, Jiaxin Zhang, Xiang Gao, Wendi Cui, Peng Li, and Kamalika Das. Gradient-guided attention map editing: Towards efficient contextual hallucination mitigation. *arXiv preprint arXiv:2503.08963*, 2025.

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

[18] 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. *arXiv preprint arXiv:2503.08342*, 2025.

[19] 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*, pages 14625–14634, 2025.

[20] Han Sun, Qin Li, Peixin Wang, and Min Zhang. Mitigating object hallucinations in lvlms via attention imbalance rectification. *arXiv preprint arXiv:2603.24058*, 2026.

[21] Xinlei Yu, Chengming Xu, Guibin Zhang, Yongbo He, Zhangquan Chen, Zhucun Xue, Jiangning Zhang, Yue Liao, Xiaobin Hu, Yu-Gang Jiang, et al. Visual multi-agent system: Mitigating hallucination snowballing via visual flow. *arXiv preprint arXiv:2509.21789*, 2025.

[22] Xingyu Zhu, Kesen Zhao, Liang Yi, Shuo Wang, Zhicai Wang, Beier Zhu, and Hanwang Zhang. Look carefully: Adaptive visual reinforcements in multimodal large language models for hallucination mitigation. *arXiv preprint arXiv:2602.24041*, 2026.

[23] Shuo Yang, Yuwei Niu, Yuyang Liu, Yang Ye, Bin Lin, and Li Yuan. Look-back: Implicit visual re-focusing in mllm reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 11694–11702, 2026.

[24] Dawei Su and Dongsheng Wang. Retllm: Training and data-free mllms for multimodal information retrieval. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 9812–9816. IEEE, 2026.

[25] Wenhui Li, Xinqi Su, Dan Song, Lanjun Wang, Kun Zhang, and An-An Liu. Towards deconfounded image-text matching with causal inference. In *Proceedings of the 31st ACM international conference on multimedia*, pages 6264–6273, 2023.

[26] Xinmiao Hu, Chun Wang, Ruihe An, ChenYu Shao, Xiaojun Ye, Sheng Zhou, and Liangcheng Li. Causal-llava: Causal disentanglement for mitigating hallucination in multimodal large language models. *arXiv preprint arXiv:2505.19474*, 2025.

[27] Jusheng Zhang, Kaitong Cai, Yijia Fan, Jian Wang, and Keze Wang. Cf-vlm: Counterfactual vision-language fine-tuning. *arXiv preprint arXiv:2506.17267*, 2025.

[28] Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Kening Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, et al. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*, 2024.

[29] Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James Glass. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. *arXiv preprint arXiv:2407.07071*, 2024.

[30] Jianjiang Yang, Yanshu Li, and Ziyan Huang. Reloop:" seeing twice and thinking backwards" via closed-loop training to mitigate hallucinations in multimodal understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 4162–4179, 2025.

[31] Shawn Li, Jiashu Qu, Yuxiao Zhou, Yuehan Qin, Tiankai Yang, and Yue Zhao. Treble counterfactual vlms: A causal approach to hallucination. *arXiv preprint arXiv:2503.06169*, 2025.

[32] Qiyan Zhao, Xiaofeng Zhang, Yiheng Li, Yun Xing, Xiaosong Yuan, Feilong Tang, Sinan Fan, Xuhang Chen, Da-Han Wang, and Xu-Yao Zhang. Mca-llava: Manhattan causal attention for reducing hallucination in large vision-language models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 3981–3990, 2025.

[33] Liu Yu, Zhonghao Chen, Ping Kuang, Zhikun Feng, Fan Zhou, Lan Wang, and Gillian Dobbie. Causally-grounded dual-path attention intervention for object hallucination mitigation in lvlms. *arXiv preprint arXiv:2511.09018*, 2025.

[34] Qiming Li, Zekai Ye, Xiaocheng Feng, Weihong Zhong, Weitao Ma, and Xiachong Feng. Causal tracing of object representations in large vision language models: Mechanistic interpretability and hallucination mitigation. *arXiv preprint arXiv:2511.05923*, 2025.

[35] Shenshen Li, Xing Xu, Wenxin Meng, Jingkuan Song, Chong Peng, and Heng Tao Shen. Mitigating hallucinations in large vision-language models via reasoning uncertainty-guided refinement. *IEEE Transactions on Multimedia*, 2025.

[36] Zhe Xu, Zhicai Wang, Junkang Wu, Jinda Lu, and Xiang Wang. Causal-halbench: Uncovering lvlms object hallucinations through causal intervention. *arXiv preprint arXiv:2511.10268*, 2025.

[37] Hao Fang, Changle Zhou, Jiawei Kong, Kuofeng Gao, Bin Chen, Tao Liang, Guojun Ma, and Shu-Tao Xia. Grounding language with vision: A conditional mutual information calibrated decoding strategy for reducing hallucinations in lvlms. *arXiv preprint arXiv:2505.19678*, 2025.

[38] Yung-Sung Chuang, Yujia Xie, Hongyin Luo, Yoon Kim, James Glass, and Pengcheng He. Dola: Decoding by contrasting layers improves factuality in large language models. *arXiv preprint arXiv:2309.03883*, 2023.

[39] Zhaorun Chen, Zhuokai Zhao, Hongyin Luo, Huaxiu Yao, Bo Li, and Jiawei Zhou. Halc: Object hallucination reduction via adaptive focal-contrast decoding. *arXiv preprint arXiv:2403.00425*, 2024.

[40] Avshalom Manevich and Reut Tsarfaty. Mitigating hallucinations in large vision-language models (lvlms) via language-contrastive decoding (lcd). *arXiv preprint arXiv:2408.04664*, 2024.

[41] Jiaming Li, Jiacheng Zhang, Zequn Jie, Lin Ma, and Guanbin Li. Mitigating hallucination for large vision language model by inter-modality correlation calibration decoding. *arXiv preprint arXiv:2501.01926*, 2025.

[42] Dayong Liang, Xiao-Yong Wei, and Changmeng Zheng. Multi-agent undercover gaming: Hallucination removal via counterfactual test for multimodal reasoning. *arXiv preprint arXiv:2511.11182*, 2025.

[43] Ren Zhibo, Wang Huizhen, Zhu Muhua, Wang Yichao, Xiao Tong, and Zhu Jingbo. Overcoming language priors with counterfactual inference for visual question answering. In *Proceedings of the 22nd Chinese National Conference on Computational Linguistics*, pages 600–610, 2023.

[44] Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, et al. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 26147–26159, 2025.

[45] Wan Zhang and Jing Zhang. Hallucination mitigation for retrieval-augmented large language models: a review. *Mathematics*, 13(5):856, 2025.

[46] Shao-Jun Xia, Huixin Zhang, Yifan Jiang, Xiaoyang Chen, Yike Chen, Zitong Li, Zhongwei Wan, and Anlan Sun. Rethinking hallucinations: A cognitive-inspired taxonomy and comprehensive survey in large language models, large vision-language models, and multimodal large language models. *Large Vision-Language Models, and Multimodal Large Language Models (November 09, 2025)*, 2025.

[47] Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*, 2024.

[48] Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. Mitigating hallucination in large vision-language models via modular attribution and intervention. In *Adaptive Foundation Models: Evolving AI for Personalized and Efficient Learning*, 2025.

[49] Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13872–13882, 2024.

[50] Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. Analyzing and mitigating object hallucination in large vision-language models. *arXiv preprint arXiv:2310.00754*, 2023.

[51] Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. *arXiv preprint arXiv:2403.18715*, 2024.

[52] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. 2023.

[53] Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Diego Rojas, Guanyu Feng, Hanlin Zhao, Hanyu Lai, Hao Yu, Hongning Wang, Jiadai Sun, Jiajie Zhang, Jiale Cheng, Jiayi Gui, Jie Tang, Jing Zhang, Juanzi Li, Lei Zhao, Lindong Wu, Lucen Zhong, Mingdao Liu, Minlie Huang, Peng Zhang, Qinkai Zheng, Rui Lu, Shuaiqi Duan, Shudan Zhang, Shulin Cao, Shuxun Yang, Weng Lam Tam, Wenyi Zhao, Xiao Liu, Xiao Xia, Xiaohan Zhang, Xiaotao Gu, Xin Lv, Xinghan Liu, Xinyi Liu, Xinyue Yang, Xixuan Song, Xunkai Zhang, Yifan An, Yifan Xu, Yilin Niu, Yuantao Yang, Yueyan Li, Yushi Bai, Yuxiao Dong, Zehan Qi, Zhaoyu Wang, Zhen Yang, Zhengxiao Du, Zhenyu Hou, and Zihan Wang. Chatglm: A family of large language models from glm-130b to glm-4 all tools. 2024.

[54] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

[55] 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*, pages 740–755. Springer, 2014.

[56] 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.

[57] 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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025.

[58] 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*, pages 216–233. Springer, 2024.

[59] Jielin Qiu, Yi Zhu, Xingjian Shi, Florian Wenzel, Zhiqiang Tang, Ding Zhao, Bo Li, and Mu Li. Benchmarking robustness of multimodal image-text models under distribution shift. *arXiv preprint arXiv:2212.08044*, 2022.

A Appendix

B More Experiment Details and Results

B.1 Datasets and Evaluation Metrics

POPE [11]: The Polling-based Object Probing Evaluation (POPE) stands as a popular benchmark in discerning hallucination at the object level. POPE employs a binary question-answering format, inquiring LVLMs to determine the presence or absence of a specified object within a given image. In this paper, we conduct experiments on subsets of MSCOCO, comprising 500 images alongside six questions per image. POPE introduces three settings within each subset: random (selecting absent

objects at random), popular (choosing the most frequently occurring objects in the dataset as absent), and adversarial (selecting absent objects that often co-occur with ground-truth objects). We adopt Accuracy and F1 score as the evaluation metrics.

Perturbed POPE: Following prior work [59], we conduct experiments on the POPE benchmark and examine distribution shifts by injecting Gaussian noise into the image inputs. Performance on the Gaussian-noise–perturbed POPE benchmark at different noise levels is reported in Tables 9–13. This evaluation considers additive Gaussian noise rather than the full range of real-world image degradations. Here, we first normalize the pixel values and then add random Gaussian noise with standard deviations of 0.08, 0.12, 0.18, 0.26, and 0.38 to represent increasing levels of severity.

MME [57]: MME benchmark serves as a comprehensive tool for assessing the capabilities of LVLMs across both perception and cognition, spanning a total of 14 tasks. Among these, tasks focusing on existence, count, position, and color are specifically designed as hallucination discrimination benchmarks. These tasks aim to scrutinize both object-level and attribute-level hallucination symptoms. MME similarly utilizes a question-answering format to facilitate this evaluation. Consequently, task scores are reported as the evaluation metric for measuring performance.

MM-Vet [56]: MM-Vet benchmark is an evaluation benchmark that examines large multimodal models (LMMs) on complicated multimodal tasks. MM-Vet defines 6 core VL capabilities (such as optical character recognition, visual recognition capability, spatial awareness, language generation, model's arithmetic capability, and knowledge-related capabilities), and examines the 16 integrations of interest derived from their combinations. For evaluation metrics, an LLM-based evaluator is used for open-ended outputs. The evaluator enables the evaluation across different question types and answer styles, resulting in a unified scoring metric.

MMBench [58]: MMBench is a systematically designed objective evaluation benchmark to robustly evaluate different abilities of large vision-language models. Currently, MMBench contains over 3000 multiple-choice questions covering 20 different ability dimensions, such as object localization and social reasoning, for evaluating vision-language models. Each ability dimension encompasses over 125 questions, with the quantity of questions per ability maintained at a roughly equal level. The distribution facilitates a balanced and thorough assessment of these abilities. In this paper, we utilized MMBench_DEV_EN for the MMBench evaluation.

Evaluation Metrics CHAIR [54] evaluates how well the generated captions align with the visual content of the given image. It consists of two variants: CHAIR_S, which measures hallucination at the sentence level, and CHAIR_I, which evaluates hallucination at the object level by comparing the number of hallucinated objects to the total number of mentioned objects. The CHAIR metric includes per-instance evaluation (CHAIR_I) and per-sentence evaluation (CHAIR_S), which are defined as follows:

$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all objects mentioned}\}|}, \tag{8}$$

$$\text{CHAIR}_S = \frac{|\{\text{sentences with hallucinated objects}\}|}{|\{\text{all sentences}\}|}. \tag{9}$$

For the image captioning task, a randomly sampled image set from the COCO Val2014 dataset was used across all experiments, with the prompt "Please describe this image in detail.". Performance was evaluated using CHAIR_S and CHAIR_I. In the MM-Vet evaluation, we used the online evaluator powered by OpenAI GPT-4.1 to assess the generated results. For the MMBench evaluation, we used an online evaluator powered by OpenAI GPT-3.5-turbo-0613.

B.2 Hyperparameter Setup

Unless otherwise specified, we employed greedy search as the default decoding strategy across all benchmark evaluations. For benchmarks MME, MMBench, and MM-Vet, questions from the annotation files were used as prompts, formatted to fit the chat templates of each respective LVLM. All parameter settings adhered to the default configurations specified in the respective papers and code repositories. Following the original papers, all MemVR tests were conducted using a greedy decoding approach, with sample=False, temperature=0, threshold=0.75, and beam=1. The intermediate layers

Table 7: Performance evaluation on POPE benchmark. We report accuracy and F1-score under three settings (Random, Popular, and Adversarial), together with their average, with absolute changes compared to the LLaVA-1.5 baseline.

Evaluation	Methods	Random		Popular		Adversarial		Average	
		Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑
MSCOCO	LLaVA-1.5-7B	88.18 ↑0.0	87.34 ↑0.0	87.30 ↑0.0	86.17 ↑0.0	**85.17** ↑0.0	84.21 ↑0.0	86.89 ↑0.0	85.91 ↑0.0
	OPERA	86.05 ↓2.13	84.73 ↓2.61	85.33 ↓1.97	83.66 ↓2.51	83.30 ↓1.87	81.80 ↓2.41	84.89 ↓2.00	83.40 ↓2.51
	ICD	**89.14** ↑0.96	**89.37** ↑2.03	85.63 ↓1.67	86.05 ↓0.12	79.37 ↓5.80	81.11 ↓3.10	84.71 ↓2.18	82.18 ↓3.73
	VCD	86.46 ↓1.72	85.22 ↓2.12	85.57 ↓1.73	83.99 ↓2.18	83.40 ↓1.77	82.03 ↓2.18	85.14 ↓1.75	83.75 ↓2.16
	MemVR	88.18 ↑0.0	87.36 ↑0.02	87.03 ↓0.27	85.94 ↓0.23	85.16 ↓0.01	84.24 ↑0.03	86.79 ↓0.10	85.85 ↓0.06
	AVED (Ours)	88.72 ↑0.54	88.08 ↑0.74	**87.50** ↑0.20	**86.60** ↑0.43	85.13 ↓0.04	**84.46** ↑0.25	**87.10** ↑0.21	**86.40** ↑0.49

Table 8: Performance evaluation on POPE benchmark. We report accuracy and F1-score under three settings (Random, Popular, and Adversarial), together with their average, with absolute changes compared to the Qwen2.5-VL baseline.

Evaluation	Methods	Random		Popular		Adversarial		Average	
		Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑
MSCOCO	Qwen2.5-VL-7B	88.56 ↑0.0	87.63 ↑0.0	87.80 ↑0.0	86.57 ↑0.0	86.70 ↑0.0	85.54 ↑0.0	87.69 ↑0.0	86.58 ↑0.0
	OPERA	88.68 ↑0.12	87.92 ↑0.29	87.83 ↑0.03	86.72 ↑0.15	86.88 ↑0.18	85.84 ↑0.30	87.80 ↑0.11	86.83 ↑0.25
	ICD	90.10 ↑1.54	89.53 ↑1.90	89.00 ↑1.20	88.19 ↑1.62	87.30 ↑0.60	86.62 ↑1.08	88.80 ↑1.11	88.11 ↑1.53
	VCD	88.73 ↑0.17	87.88 ↑0.25	87.80 ↑0.0	86.62 ↑0.05	86.87 ↑0.17	85.83 ↑0.29	87.80 ↑0.11	86.78 ↑0.20
	MemVR	89.04 ↑0.48	88.20 ↑0.57	87.93 ↑0.13	86.74 ↑0.17	87.07 ↑0.37	85.97 ↑0.43	88.01 ↑0.32	86.97 ↑0.39
	Ours	**90.79** ↑2.23	**90.34** ↑2.71	**89.47** ↑1.67	**88.80** ↑2.23	**87.93** ↑1.23	**87.38** ↑1.84	**89.39** ↑1.70	**88.84** ↑2.26

Table 9: Performance evaluation on Perturbed POPE benchmark. The perturbation level is 0.08. We report accuracy and F1-score under three settings (Random, Popular, and Adversarial), together with their average, with absolute changes compared to the Qwen2.5-VL-7B baseline. Green and red arrows indicate performance gains and drops.

Evaluation	Methods	Random		Popular		Adversarial		Average	
		Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑
MSCOCO	Qwen2.5-VL-7B	86.84 ↑0.0	85.49 ↑0.0	86.03 ↑0.0	84.34 ↑0.0	84.90 ↑0.0	83.28 ↑0.0	85.92 ↑0.0	84.37 ↑0.0
	OPERA	86.98 ↑0.14	85.65 ↑0.16	86.11 ↑0.08	84.46 ↑0.12	84.87 ↓0.03	83.26 ↓0.02	85.99 ↑0.07	84.46 ↑0.09
	ICD	**89.76** ↑2.92	**89.16** ↑3.67	87.77 ↑1.74	86.96 ↑2.62	85.70 ↑0.80	85.14 ↑1.86	87.74 ↑1.82	87.09 ↑2.72
	VCD	86.74 ↓0.10	85.40 ↓0.09	86.20 ↑0.17	84.66 ↑0.32	84.87 ↓0.03	83.44 ↑0.16	85.93 ↑0.01	84.50 ↑0.13
	MemVR	87.04 ↑0.20	85.78 ↑0.29	86.20 ↑0.17	84.55 ↑0.21	85.17 ↑0.27	83.69 ↑0.41	86.14 ↑0.22	84.67 ↑0.30
	Ours	89.48 ↑2.64	88.80 ↑3.31	**87.97** ↑1.94	**87.04** ↑2.70	**86.43** ↑1.53	**85.64** ↑2.36	**87.96** ↑2.04	**87.16** ↑2.79

are selected based on the model depth (e.g., layers 5–16 for LLaVA-1.5 and layers 9–16 for Qwen2.5-VL). The retracing ratio is set to 0.12 for LLaVA-1.5 and 0.28 for Qwen2.5-VL, and λ is fixed to a constant value (e.g., 0.75). For VCD tests, we set sample=False, temperature=0, noise step=500, and the plausibility constraint hyperparameter λ to 0.1, while α, which controls the degree of contrastive emphasis, was set to 1, following the default parameter settings from the original code and literature. OPERA tests were configured with beam=5, sample=True, scale factor=50, threshold=15, num-attn-candidates=5, and penalty weights=1.

As for our AVED based on LLaVA-1.5, VAC is applied to the hallucination-prone layers identified in prior work [48], where the attention heads 14-20 are calibrated according to Algorithm 1, with an attention calibration strength of $\alpha = 1.0$ for the image captioning task and $\alpha = 0.45$ for other visual question answering benchmarks. LLR is applied to higher decoding layers ranging from layer 21 to 32, using a top-K value of 15 and a logits reweighting coefficient $\beta = 2.0$. For Qwen2.5-VL and Qwen3.5-4B, LLR is applied to layers 14-28, with the top-K value of 15 and a logits reweighting coefficient $\beta = 1.0$. For GLM-4V-9B, LLR is applied to layers 30-40, with the top-K value of 15 and a logits reweighting coefficient $\beta = 1.0$.

B.3 More Experiment Results

Ablation Study Results and Selection Principles on Hyperparameters We conducted sensitivity analyses for α, β, K, and layer ranges using LLaVA-1.5-7B on the COCO dataset. The key observations and selection principles are summarized below:

Table 10: Performance evaluation on Perturbed POPE benchmark. The perturbation level is 0.12. We report accuracy and F1-score under three settings (Random, Popular, and Adversarial), together with their average, with absolute changes compared to the Qwen2.5-VL-7B baseline. Green and red arrows indicate performance gains and drops.

Evaluation	Methods	Random		Popular		Adversarial		Average	
		Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑
MSCOCO	Qwen2.5-VL-7B	85.60 ↑0.0	83.90 ↑0.0	84.83 ↑0.0	82.76 ↑0.0	83.60 ↑0.0	81.61 ↑0.0	84.68 ↑0.0	82.76 ↑0.0
	OPERA	86.29 ↑0.69	84.82 ↑0.92	85.43 ↑0.60	83.61 ↑0.85	83.90 ↑0.30	82.20 ↑0.59	85.21 ↑0.53	83.54 ↑0.78
	ICD	88.20 ↑2.60	**87.56** ↑3.66	85.63 ↑0.80	84.88 ↑2.12	83.63 ↑0.03	83.17 ↑1.56	85.83 ↑1.15	85.20 ↑2.44
	VCD	85.91 ↑0.31	84.35 ↑0.45	84.87 ↑0.04	82.93 ↑0.17	83.20 ↓0.40	81.37 ↓0.24	84.66 ↓0.02	82.89 ↑0.13
	MemVR	86.12 ↑0.52	84.57 ↑0.67	85.20 ↑0.37	83.30 ↑0.54	83.87 ↑0.27	82.06 ↑0.45	85.06 ↑0.38	83.31 ↑0.55
	Ours	**88.42** ↑2.82	87.53 ↑3.63	**86.73** ↑1.90	**85.60** ↑2.84	**85.00** ↑1.40	**84.02** ↑2.41	**86.92** ↑2.24	**86.72** ↑3.96

Table 11: Performance evaluation on Perturbed POPE benchmark. The perturbation level is 0.18. We report accuracy and F1-score under three settings (Random, Popular, and Adversarial), together with their average, with absolute changes compared to the Qwen2.5-VL-7B baseline. Green and red arrows indicate performance gains and drops.

Evaluation	Methods	Random		Popular		Adversarial		Average	
		Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑
MSCOCO	Qwen2.5-VL-7B	83.23 ↑0.0	80.86 ↑0.0	82.43 ↑0.0	79.64 ↑0.0	81.43 ↑0.0	78.73 ↑0.0	82.37 ↑0.0	79.75 ↑0.0
	OPERA	83.54 ↑0.31	81.34 ↑0.48	82.60 ↑0.17	80.00 ↑0.36	81.50 ↑0.07	79.00 ↑0.27	82.55 ↑0.18	80.11 ↑0.36
	ICD	**87.76** ↑4.53	**87.20** ↑6.34	85.07 ↑2.64	**84.33** ↑4.69	81.50 ↑0.07	81.17 ↑2.44	84.78 ↑2.41	84.24 ↑4.49
	VCD	83.71 ↑0.48	81.53 ↑0.67	82.53 ↑0.10	79.98 ↑0.34	81.40 ↓0.03	79.05 ↑0.32	82.55 ↑0.18	80.19 ↑0.44
	MemVR	84.71 ↑1.48	82.79 ↑1.93	83.50 ↑1.07	81.21 ↑1.57	82.37 ↑0.94	80.18 ↑1.45	83.53 ↑1.16	81.39 ↑1.64
	Ours	87.63 ↑4.40	86.64 ↑5.78	**85.37** ↑2.94	84.17 ↑4.53	**83.07** ↑1.64	**82.13** ↑3.40	**85.36** ↑2.99	**84.31** ↑4.56

Table 12: Performance evaluation on Perturbed POPE benchmark. The perturbation level is 0.26. We report accuracy and F1-score under three settings (Random, Popular, and Adversarial), together with their average, with absolute changes compared to the Qwen2.5-VL-7B baseline. Green and red arrows indicate performance gains and drops.

Evaluation	Methods	Random		Popular		Adversarial		Average	
		Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑
MSCOCO	Qwen2.5-VL-7B	81.37 ↑0.0	78.27 ↑0.0	79.93 ↑0.0	76.43 ↑0.0	78.90 ↑0.0	75.51 ↑0.0	80.07 ↑0.0	76.74 ↑0.0
	OPERA	81.72 ↑0.35	78.80 ↑0.53	80.47 ↑0.54	77.15 ↑0.72	79.23 ↑0.33	76.05 ↑0.54	80.47 ↑0.40	76.05 ↓0.69
	ICD	**86.01** ↑4.64	**85.67** ↑7.40	81.83 ↑1.90	81.63 ↑5.20	78.37 ↓0.53	78.74 ↑3.23	82.07 ↑2.00	82.01 ↑5.27
	VCD	82.10 ↑0.73	79.53 ↑1.26	80.17 ↑0.24	77.11 ↑0.68	78.70 ↓0.20	75.90 ↑0.39	80.32 ↑0.25	77.51 ↑0.77
	MemVR	82.16 ↑0.79	79.49 ↑1.22	80.30 ↑0.37	77.30 ↑0.87	79.20 ↑0.30	76.33 ↑0.82	80.55 ↑0.48	77.71 ↑0.97
	Ours	85.78 ↑4.41	84.69 ↑6.42	**83.10** ↑3.17	**81.87** ↑5.44	**80.67** ↑1.77	**79.79** ↑4.28	**83.18** ↑3.11	**82.12** ↑5.38

Table 13: Performance evaluation on Perturbed POPE benchmark. The perturbation level is 0.38. We report accuracy and F1-score under three settings (Random, Popular, and Adversarial), together with their average, with absolute changes compared to the Qwen2.5-VL-7B baseline. Green and red arrows indicate performance gains and drops.

Evaluation	Methods	Random		Popular		Adversarial		Average	
		Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑	Accuracy ↑	F1-score ↑
MSCOCO	Qwen2.5-VL-7B	77.84 ↑0.0	73.49 ↑0.0	75.40 ↑0.0	70.78 ↑0.0	74.43 ↑0.0	69.98 ↑0.0	75.89 ↑0.0	71.42 ↑0.0
	OPERA	78.04 ↑0.20	73.87 ↑0.38	75.60 ↑0.20	71.16 ↑0.38	74.33 ↓0.10	70.11 ↑0.13	75.99 ↑0.10	71.71 ↑0.29
	ICD	81.92 ↑4.08	80.58 ↑7.09	75.30 ↓0.10	74.65 ↑3.87	75.13 ↑0.70	72.18 ↑2.20	77.45 ↑1.56	75.80 ↑4.38
	VCD	78.69 ↑0.85	75.22 ↑1.73	75.83 ↑0.43	72.32 ↑1.54	74.33 ↓0.10	71.23 ↑1.25	76.29 ↑0.40	72.92 ↑1.50
	MemVR	78.93 ↑1.09	75.07 ↑1.58	76.17 ↑0.77	72.08 ↑1.30	75.07 ↑0.64	71.16 ↑1.18	76.72 ↑0.83	72.77 ↑1.35
	Ours	**83.99** ↑6.15	**83.15** ↑9.66	**79.10** ↑3.70	**78.58** ↑7.80	**76.50** ↑2.07	**76.54** ↑6.56	**79.83** ↑3.94	**79.42** ↑8.00

Table 14: Ablation study results on hyperparameter α.

α	**1.0**	0.75	0.5	0.25	0.00
$CHAIR_S$	**46.80**	52.83	51.00	51.00	51.40
$CHAIR_I$	**13.14**	14.70	14.72	14.72	14.32

Table 15: Ablation study results on hyperparameter β.

β	0.1	0.5	1.0	1.5	**2.0**	2.50	3.0	5.0
CHAIR_S	46.80	47.0	47.20	47.20	**46.80**	47.60	48.60	50.30
CHAIR_I	13.80	13.87	13.77	14.12	**13.14**	13.76	13.93	14.13

Table 16: Ablation study results on layer range in LLR.

Layer Range	**21–32**	16–32	26–32	16–26	21–26
CHAIR_S	**46.80**	47.20	48.90	48.20	47.60
CHAIR_I	**13.14**	13.15	14.72	14.23	13.25

Table 17: Ablation study results on hyperparameter K.

K	5	10	15	20	30	100	200	All
CHAIR_S	48.80	48.00	**46.80**	47.20	48.20	50.80	51.40	51.00
CHAIR_I	13.48	13.48	**13.14**	13.49	13.60	13.84	13.92	13.60

Table 18: Ablation of the VQA sequence representation. Accuracy and F1-score are reported on clean and perturbed POPE. Levels 0–5 correspond to Gaussian-noise standard deviations of 0, 0.08, 0.12, 0.18, 0.26, and 0.38, respectively.

Representation	0		1		2		3		4		5	
	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1	Acc	F1
Qwen2.5-VL-7B	87.7	86.6	85.9	84.4	84.7	82.8	82.4	79.8	80.1	76.7	75.9	71.4
Last token	89.0	88.0	87.2	85.5	86.3	85.2	84.8	82.4	**83.5**	80.7	**80.8**	77.7
All-text mean	88.9	87.9	86.3	86.2	86.4	85.4	84.7	82.4	83.4	80.5	79.7	77.9
Question–answer mean	**89.4**	**88.8**	**88.0**	**87.2**	**86.9**	**86.7**	**85.4**	**84.3**	83.2	**82.1**	79.8	**79.4**

- K modulates the precision of visual grounding. Empirically, as shown in Table 17, performance peaks at $K \in [5, 20]$. Larger values include less relevant visual tokens and degrade the concentration signal in the tested setting.
- α balances the original and calibrated attention in VAC. The best result is observed at $\alpha = 1.0$ (CHAIR_S: 51.40 $\rightarrow$ 46.80) in Table 14, indicating that full redistribution performs best among the tested mixing strengths in this setting.
- β balances the contribution of intermediate layers in LLR. Table 15 shows similar CHAIR values for $\beta \in [0.1, 2.5]$ in the tested setting.
- Layer Range: In Table 16, the 21–32 range gives the best result among the tested LLaVA-1.5 ranges. This supports using later layers as intervention points in this setting.
- Adaptation Strategy: The reported experiments evaluate model-specific configurations on several LVLMs, including Qwen3.5-4B and Qwen2.5-VL-7B (see results in Table 4 and Table 22). For new architectures, we recommend a set of initial hyperparameter settings: K should be set to 10%–20% of the visual sequence length, $\alpha \in [0.5, 1.0]$, and $\beta \in \{0.1, 1.0, 2.0\}$, while LLR should target approximately the final 1/3 of the model layers. These settings should be validated for each new architecture; the reported experiments do not establish a universal configuration.

Attention-guided top-K selection. Table 19 directly tests whether VEC should use a selective evidence set. Selecting $K = 15$ attended visual tokens achieves $\text{CHAIR}_S/\text{CHAIR}_I = 46.80/13.14$, whereas including all visual tokens degrades the result to $51.00/13.60$. The deterioration for very large K confirms that background and weakly relevant tokens dilute the cross-modal consistency signal. This ablation supports the top-K evidence filtering used near Equation 5, rather than treating

Table 20: Evaluation on the CHAIR benchmark based on LLaVA-1.5-7B. CHAIR_S and CHAIR_I (lower is better) measure hallucinations. Recall (higher is better) measures the recall rate of the correct object. The maximum generation length is 1024 tokens.

Method	$\text{CHAIR}_S \downarrow$	$\text{CHAIR}_I \downarrow$	Recall ↑
Greedy	52.80	13.94	77.31
AD-HH	34.60	9.12	71.19
AVED (Ours)	46.80	13.14	**77.88**
AD-HH+Ours	**32.80**	**7.39**	71.67

Table 21: Comparison with Text Attention Calibration based on LLaVA-1.5-7B. We report absolute performance as well as absolute changes from the baseline.

Evaluation	Methods	MME-Hall	Object-Level		Attribute-Level		MME-Overall	MM-Vet	MMBench
		Total ↑	Existence ↑	Count ↑	Position ↑	Color ↑	Total ↑	Score ↑	Accuracy ↑
LLaVA-1.5	Baseline	643.3 ↑0.0	190.0 ↑0.0	150.0 ↑0.0	128.3 ↑0.0	170.0 ↑0.0	1864.7 ↑0.0	31.10 ↑0.0	62.80 ↑0.0
	AD-HH	648.3 ↑5.0	190.0 ↑0.0	155.0 ↑5.0	128.3 ↑0.0	170.0 ↑0.0	1880.1 ↑15.4	31.60 ↑0.5	64.94 ↑2.1
	AVED (Ours)	**660.0** ↑16.7	190.0 ↑0.0	**160.0** ↑10.0	**140.0** ↑11.7	**170.0** ↑0.0	**1900.4** ↑35.7	**32.30** ↑1.2	**65.10** ↑2.3

Table 22: Performance comparison on Qwen3.5-4B across multiple benchmarks. We report absolute performance together with absolute changes from the baseline. Best results are highlighted in bold.

Model	Method	MMBench	MM-Vet	MME	POPE-F1	POPE-Acc	Perturbed POPE-F1	Perturbed POPE-Acc
Qwen3.5-4B	Baseline	61.51 ↑0.0	73.1 ↑0.0	2320.81 ↑0.0	88.90 ↑0.0	89.50 ↑0.0	78.60 ↑0.0	81.20 ↑0.0
	MemVR	61.51 ↑0.0	72.9 ↓0.2	**2331.32** ↑10.51	89.00 ↑0.10	89.50 ↑0.0	78.70 ↑0.10	81.40 ↑0.20
	AVED (Ours)	**64.78** ↑3.27	**73.3** ↑0.2	2325.17 ↑4.36	**89.80** ↑0.90	**90.00** ↑0.50	**82.90** ↑4.30	**83.70** ↑2.50

K merely as an implementation hyperparameter. We further conduct sensitivity analyses on α, β, K, and layer ranges. Detailed results and hyperparameter selection guidelines are provided in Appendix B.3.

Detailed Experiment Results under Distribution Shift We report performance on the Gaussian-noise–perturbed POPE benchmarks across different noise levels in Tables 9-13. In these tested perturbation settings, our method has lower reported hallucination scores, including at the highest severity (0.38).

Table 19: Effect of attention-guided top-K visual-token selection on CHAIR.

K	5	10	15	20	100	200	All
CHAIR_S	48.80	48.00	**46.80**	47.20	50.80	51.40	51.00
CHAIR_I	13.48	13.48	**13.14**	13.49	13.84	13.92	13.60

Comparison with Text Attention Calibration We further compare our AVED method with the prior approach, AD-HH [48], which calibrates text attention in hallucination-prone layers. AD-HH categorizes attention heads into two groups based on their contrastive influence scores: hallucination heads, which exert a strong influence on hallucinated words, and non-hallucination heads, which have minimal impact on hallucinated words. AD-HH then suppresses hallucination by setting the text attention weights of hallucination heads to zero.

We report performance comparisons on CHAIR, MME, MM-Vet, and MMBench in Tables 20-21. As shown in Table 20, AD-HH reduces object hallucination at the cost of a decreased recall rate for correct objects. When implementing the proposed AVED on top of AD-HH, we obtain lower CHAIR_I and CHAIR_S scores in the reported setting. Moreover, as reported in Table 21, our method improves over AD-HH on the reported MME measures for object-level and attribute-level hallucinations.

B.4 Case Study

This case study aims to evaluate and present various benchmark cases across multiple domains systematically.

Table 23: Comparison of generation speed on the image captioning task (maximum generation length: 1024 tokens).

Model	Speed (token/s)	GPU Memory (Mb)
Baseline	11.6	17844
MemVR	11.3	17988
ICD	6.5	18010
VCD	5.9	17916
w / VAC	7.8	17870
w / LLR	10.2	17856

Case Study on Image Captioning We present several examples of the image captioning task in Figures 5-7. We also illustrate the intermediate-layer logits and the reweighting process of our LLR module. The results demonstrate that LLR effectively suppresses hallucinated words while promoting vision-grounded words. By leveraging intermediate-layer outputs, LLR adaptively reweights the logits of the middle layers based on their confidence, guiding the final prediction toward visually grounded semantics.

Case Study on MMbench We present several examples on MMBench in Figures 8-10. The examples show that the proposed Visual Attention Calibration (VAC) module produces attention maps that are qualitatively more aligned with the displayed text query in these cases.

Case Study on POPE under Distribution Shift We visualize the effect of Layer-wise Logits Reweighting (LLR) on the perturbed POPE benchmark in Figure 11. LLR suppresses incorrect predictions from higher layers and guides the final output to align with visual facts.

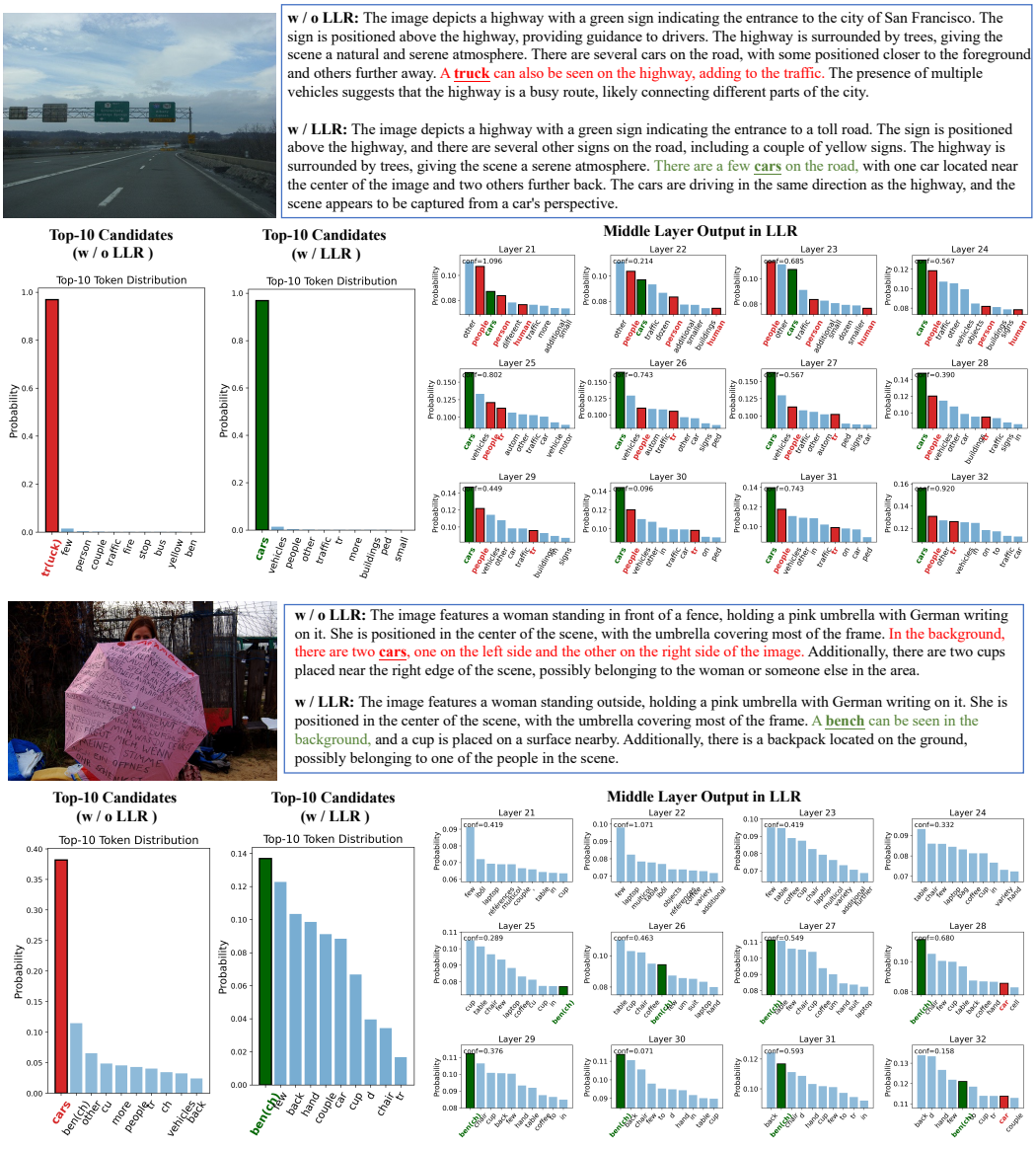

Figure 5: The effect of Layer-wise Logits Reweighting. For the first example in the figure, the proposed LLR module effectively suppresses the hallucinated words "tr(uck)" and "people", while successfully promoting the vision-grounded word "cars". For the second example in the figure, the proposed LLR module effectively suppresses the hallucinated words "cars", while successfully promoting the vision-grounded word "ben(ch)". Here, "tr(uck)" denotes that the token generated at the current step is "tr", and the subsequent token is "uck". The notation "ben(ch)" is interpreted in the same manner.

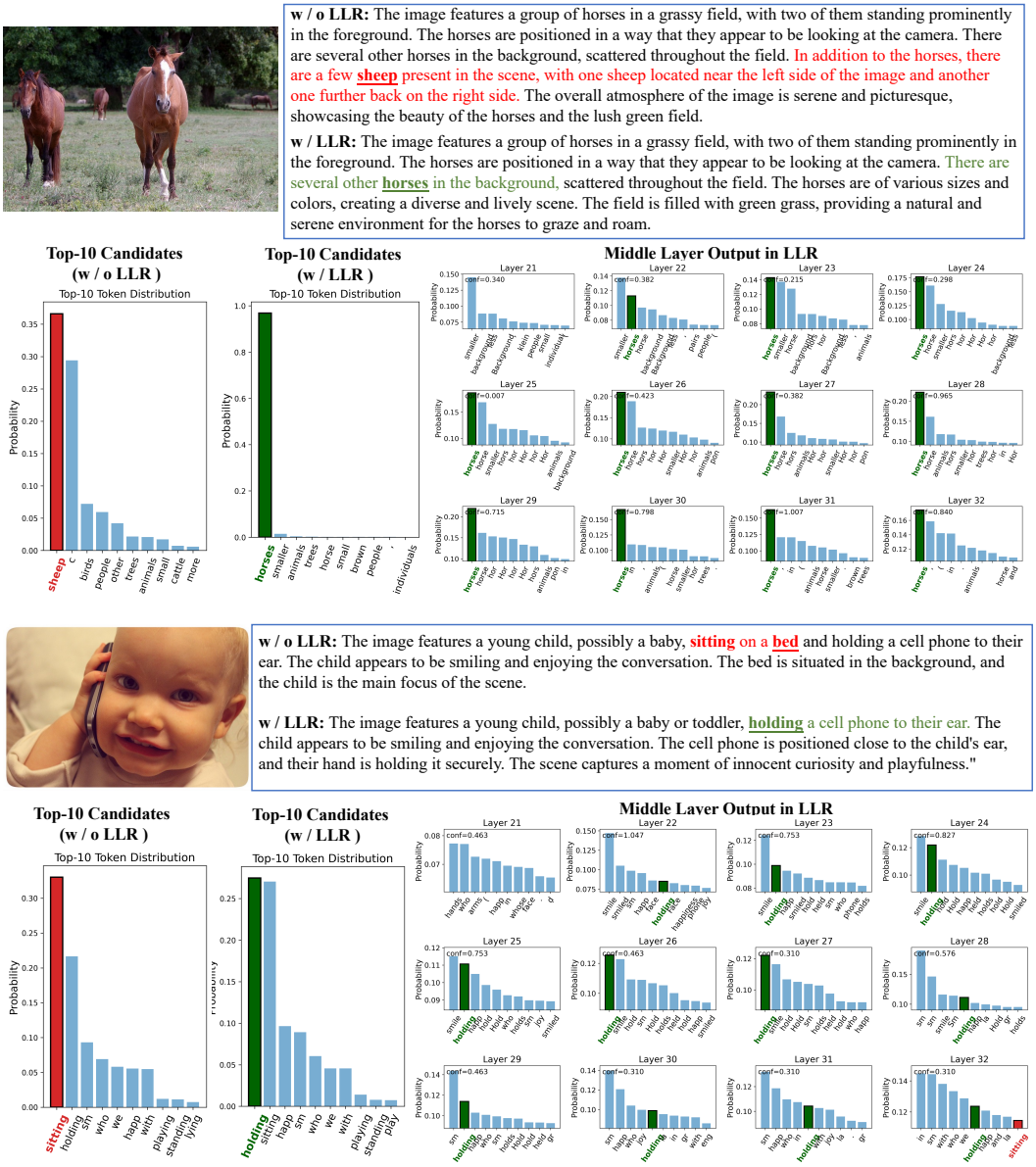

Figure 6: The effect of Layer-wise Logits Reweighting. For the first example in the figure, the proposed LLR module effectively suppresses the hallucinated words "sheep", while successfully promoting the vision-grounded word "horses". For the second example in the figure, the proposed LLR module effectively suppresses the hallucinated words "sitting", while successfully promoting the vision-grounded word "holding".

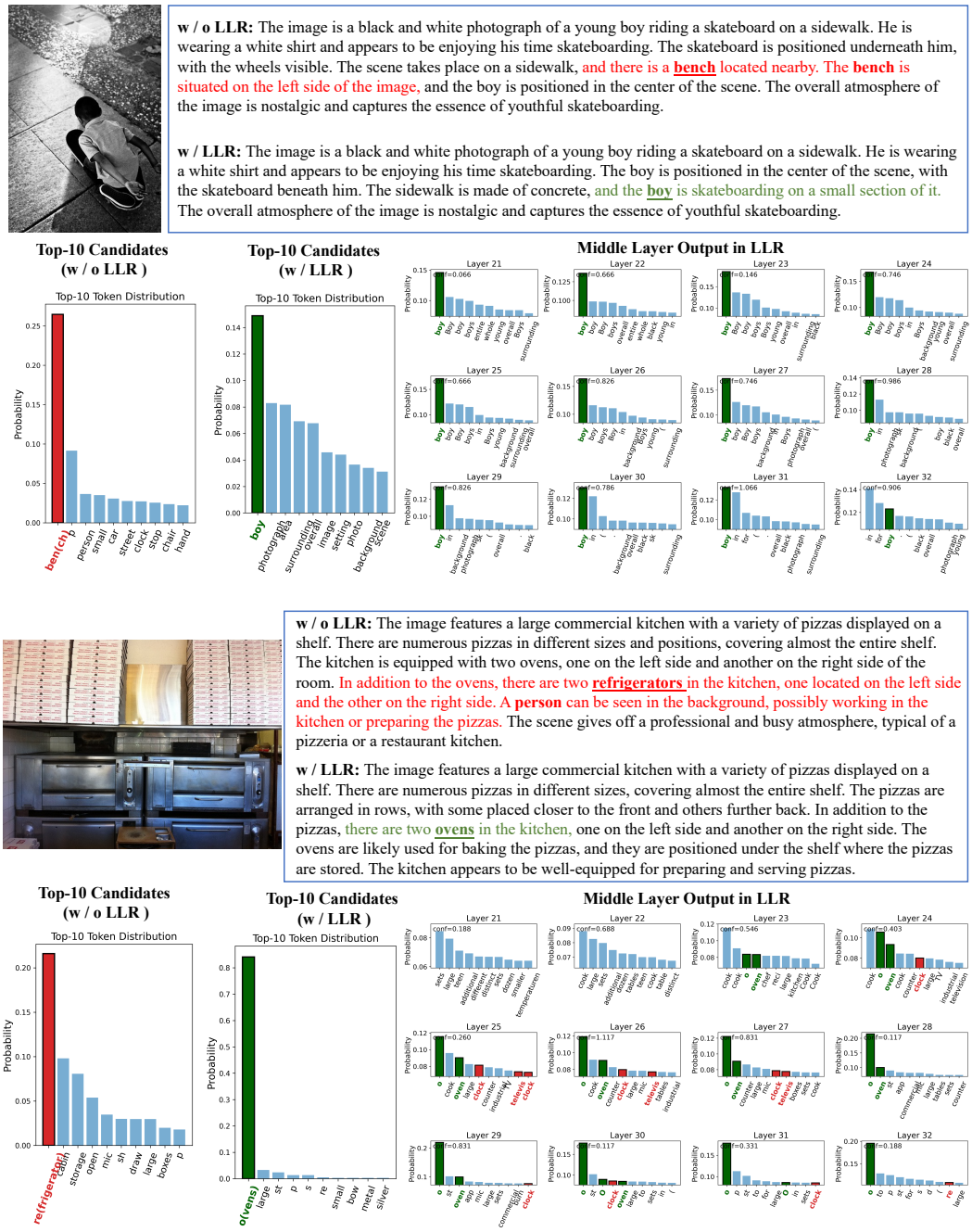

Figure 7: The effect of Layer-wise Logits Reweighting. For the first example in the figure, the proposed LLR module effectively suppresses the hallucinated words "ben(ch)", while successfully promoting the vision-grounded word "boy". For the second example in the figure, the proposed LLR module effectively suppresses the hallucinated words "re(frigerator)", "clock" and "people", while successfully promoting the vision-grounded word "o(vens)". Here, "re(frigerator)" denotes that the token generated at the current step is "re", and the subsequent token is "frigerator". The notation "ben(ch)" and "o(vens)" are interpreted in the same manner.

Question: What kind of human behavior does this picture describe?
A. A family is gardening in their backyard, planting seeds, pulling weeds, and nurturing plants while cultivating a sense of pride and accomplishment.
B. An artist is creating a sculpture in a workshop, chiseling or casting materials to form a three-dimensional work of art.
C. A man is fishing on a riverbank, casting his line and waiting patiently for a bite while enjoying the tranquility of nature.
D. A little boy and a little girl are leaning on a tree branch reading a book.

Calibrated Text-to-Image Attention:



Answer: D

Question: What kind of human behavior does this picture describe?
A. A man showed the camera his left hand, a silver ring on his ring finger, and "leave me" written in the palm of his hand with a black pen.
B. A family is picnicking in a park, enjoying food and drinks while lounging on blankets and playing games together.
C. A group of artists are painting a mural on a city wall, using brushes and spray cans to create a colorful and vibrant work of art.
D. A man is practicing his golf swing on a driving range, hitting balls with precision while aiming for distant targets.

Calibrated Text-to-Image Attention:



Answer: A

Question: What kind of human behavior does this picture describe?
A. A musician is recording music in a studio, singing or playing instruments and perfecting tracks for an upcoming album.
B. A man is playing basketball at a court, dribbling, passing, and shooting hoops with precision and skill.
C. A man shouts loudly with open arms in the rain, celebrating his regaining his freedom.
D. A group of friends are rock climbing on a cliff face, using ropes and harnesses to ascend steep heights while feeling exhilarated by the challenge.

Calibrated Text-to-Image Attention:



Answer: C

Question: What kind of human behavior does this picture What kind of human behavior does this picture describe?
A. A woman is running errands in a crowded city, navigating busy streets and bustling sidewalks to get things done efficiently.
B. A group of coworkers are taking a coffee break, chatting and socializing while recharging their energy levels.
C. A father and his son sat fishing in the lake, both tilting their heads sideways, as if they had waited a long time for the fish to be hooked.
D. A musician is performing on stage in front of a live audience, singing and playing instruments with passion and energy.

Calibrated Text-to-Image Attention:



Answer: C

Figure 8: The effect of Visual Attention Calibration. AVED integrates a Visual Attention Calibration to dynamically mitigate attention degradation on visual tokens and reallocate the attention distribution to align with the text query in these examples.

Question: Which one is the correct caption of this image?
A. An older orange van is parked next to a modern mini van in front of a small shop.
B. A black kitten laying down next to two remote controls.
C. A woman is cutting up a block of spam.
D. A man standing near the home plate swinging a bat

Calibrated Text-to-Image Attention:



Answer: B

Question: Which one is the correct caption of this image?
A. a large food truck is parked on the side of the street
B. Neither one of these people had a good flight.
C. People in a horse drawn buggy on a city street.
D. A fire hydrant with a pair of eye stickers making a face on it.

Calibrated Text-to-Image Attention:



Answer: D

Question: who is this person?
A. Ed Sheeran
B. Harry Styles
C. Idris Elba
D. Benedict Cumberbatch

Calibrated Text-to-Image Attention:



Answer: A

Question: What does this picture want to express?
A. We are expected to care for the earth.
B. We are expected to stay positive.
C. We are expected to work hard.
D. We are expected to care for green plants.

Calibrated Text-to-Image Attention:



Answer: A

Figure 9: The effect of Visual Attention Calibration. AVED integrates a Visual Attention Calibration to dynamically mitigate attention degradation on visual tokens and reallocate the attention distribution to align with the text query in these examples.

B.5 Computational Efficiency

We evaluate the computational efficiency of our method on the image captioning task. As shown in Table 23, the baseline model achieves a generation speed of 11.6 tokens/s. Existing multi-pass contrastive decoding methods, such as ICD and VCD, incur additional computational overhead, resulting in significantly lower inference speeds of 6.5 and 5.9 tokens/s, respectively.

In comparison, single-pass methods demonstrate a more favorable efficiency profile. MemVR achieves a speed of 11.3 tokens/s, while our LLR maintains competitive performance at 10.2 tokens/s. The configuration with VAC achieves 7.8 tokens/s, which still remains considerably more efficient than multi-pass contrastive decoding approaches.

The configurations with VAC and with LLR have GPU memory footprints close to the baseline. Table 23 reports component-level measurements; it does not separately report the combined VAC+LLR configuration.

C Limitations

Despite its effectiveness and efficiency advantages in avoiding multi-pass inference (e.g., ICD and VCD) and external visual re-injection, our method introduces a moderate inference overhead due to the additional computation involved in Visual Attention Calibration (VAC). As shown in Table 23, our method maintains comparable GPU memory usage to the baseline and prior methods, while achieving a generation speed of 7.8 tokens/s when VAC is applied.

This additional latency mainly stems from the attention recalibration process, which explicitly models vision-language alignment and performs attention redistribution during decoding. This design avoids a separate visual re-injection pathway but introduces additional attention computation. The configuration with LLR achieves 10.2 tokens/s.

Overall, the main trade-off of our method lies in balancing inference efficiency and calibration precision, with the reported component configurations running faster than the compared multi-pass contrastive decoding methods. This suggests a promising direction for future work to explore more lightweight or approximate attention calibration strategies to further improve runtime efficiency while preserving visual grounding effectiveness.

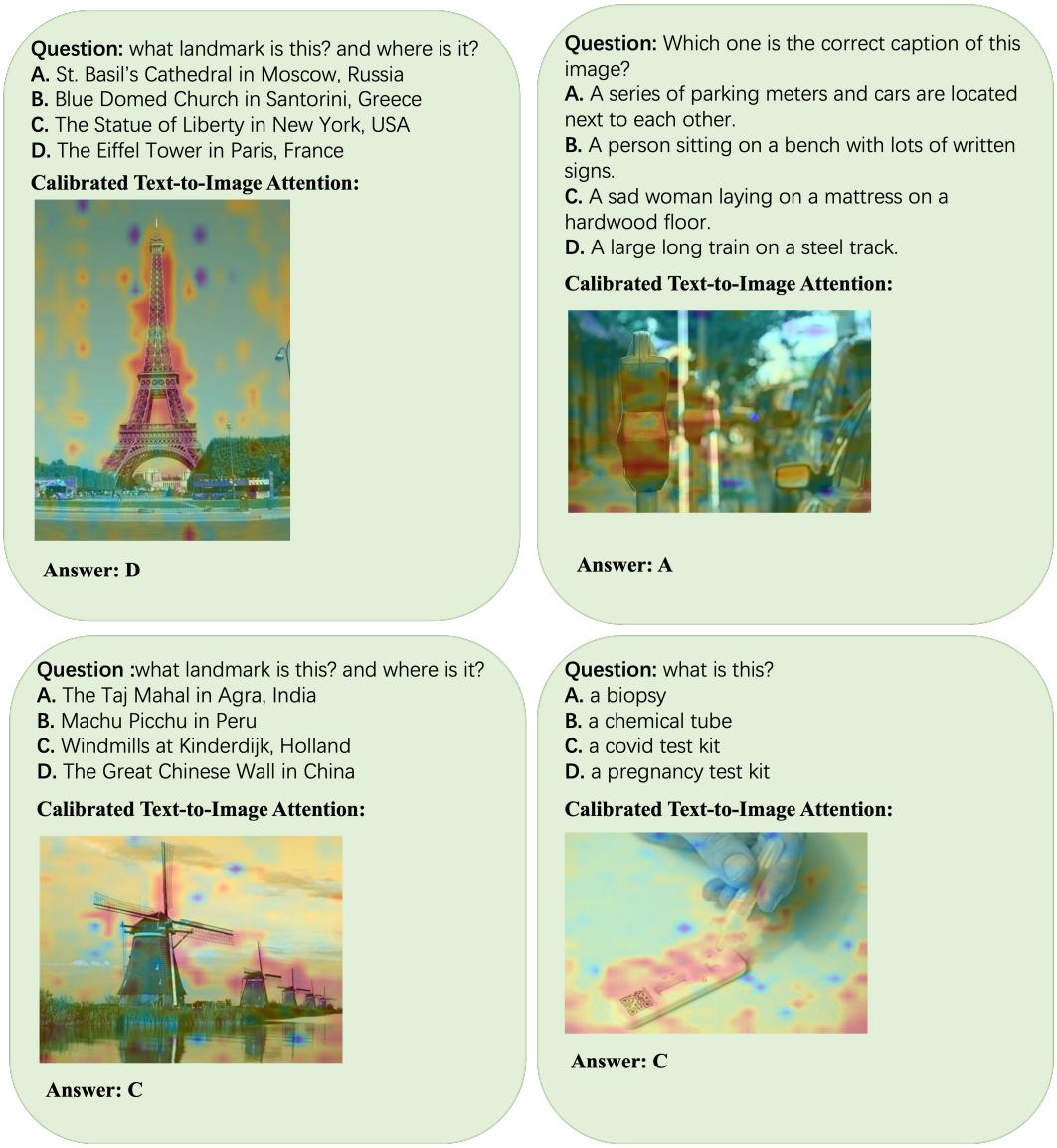

Figure 10: The effect of Visual Attention Calibration. AVED integrates a Visual Attention Calibration to dynamically mitigate attention degradation on visual tokens and reallocate the attention distribution to align with the text query in these examples.

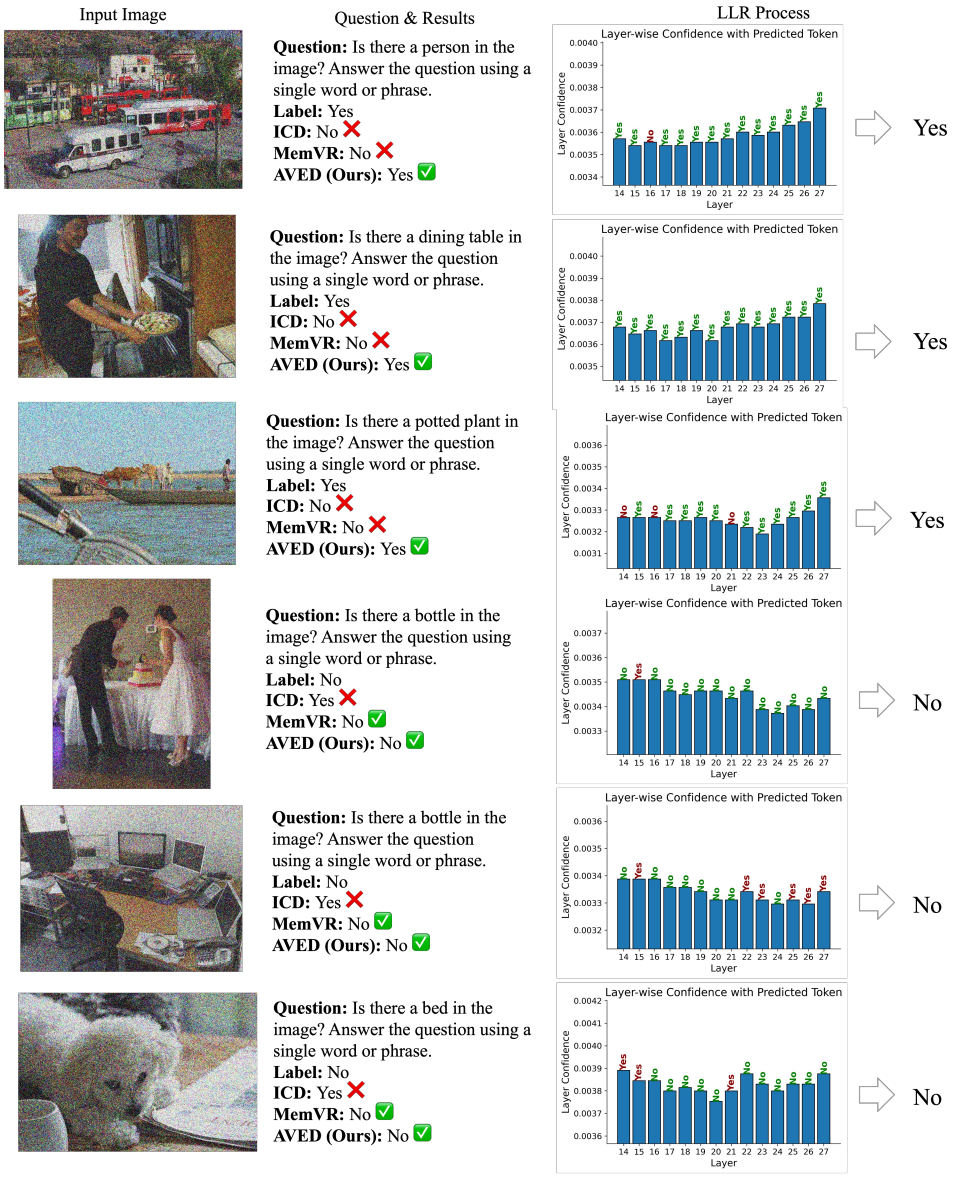

Figure 11: The effect of Layer-wise Logits Reweighting on the Perturbed POPE benchmark. The Gaussian-noise standard deviation is 0.38. By exploiting intermediate-layer outputs, LLR adaptively reweights the output logits of the middle layers based on their confidence, thereby guiding the final prediction toward visually grounded semantics.