

GLVAC: Mitigating Object Hallucination in MLLMs with Global-to-Local Visual Attention Correction

Xingxing Huang¹, Zhiguo Yang¹, Jiacheng Pi¹, Dongsheng Xu¹,
Ruizhi Zhong¹, Peipei Xu^{2,*}, Wenjie Ruan^{1,*}

¹School of Computer Science, USTC, China; ²University of Dundee, UK

Abstract

Multimodal Large Language Models (MLLMs) have achieved impressive progress in vision-language understanding, yet they remain susceptible to *object hallucination*, which refers to generating descriptions of objects that do not exist in the input image. Through comprehensive analysis, we uncover a fundamental trade-off in MLLMs: strengthening visual attention enhances grounding completeness but also increases hallucinations, especially *additive hallucinations* that appear in the trailing part of the generated sentence. To alleviate this trade-off, we propose **GLVAC**, **G**lobal-to-**L**ocal Visual Attention **C**orrection, a training-free dual-path inference framework that adaptively combines global conservatism and local enhancement in visual attention. The global path maintains stable and grounded predictions under moderate visual attention, while the local correction branch is dynamically activated for high-uncertainty tokens identified via entropy-based estimation. A confidence-guided gating mechanism then determines whether to adopt the correction output, ensuring safe and reliable refinement. GLVAC can be seamlessly integrated into existing MLLMs without retraining or external modules, effectively reducing object hallucinations while maintaining perceptual completeness across diverse benchmarks and model families. Our code is available on <https://github.com/Trusted-LLM/GLVAC>.

1 Introduction

Multimodal large language models (MLLMs) have emerged as a powerful paradigm for vision-language understanding (Li et al., 2023a; Ye et al., 2023; Liu et al., 2024c), integrating visual encoders with large language models to perform tasks such as image captioning (Hosain et al., 2019; Bucciarelli et al., 2024),

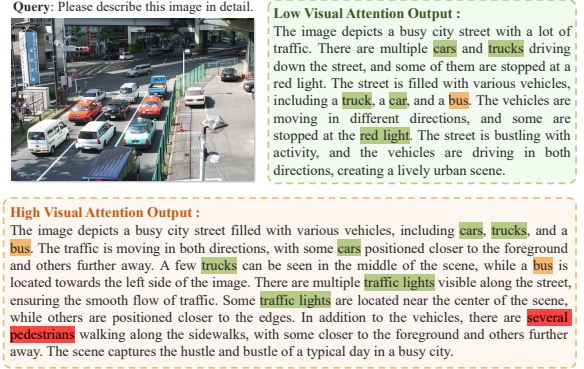


Figure 1: Comparison of model-generated descriptions under different levels of visual attention. Higher visual attention encourages more detailed descriptions but may introduce additive hallucinated objects (e.g., “pedestrians”). ■ Green marks ground-truth objects, ■ Red marks additive hallucinated objects that emerge under heightened visual attention, and ■ Orange marks other hallucinated objects.

visual question answering (Xu et al., 2024; Kuang et al., 2025), and multimodal dialogue (Huang et al., 2025; Liu et al., 2024a). Despite their strong performance, MLLMs still suffer from hallucination, where generated text is inconsistent with the visual input. Among different types of hallucinations, object hallucination is particularly concerning (Rohrbach et al., 2018), as it describes semantically plausible objects that are absent from the input image (Gunjal et al., 2024; Liu et al., 2024b). Such discrepancies undermine reliability and pose risks in high-stakes applications, including medical diagnosis (Liu et al., 2023; Zhou et al., 2024) and autonomous driving (Cui et al., 2024; Wang et al., 2023b). Therefore, ensuring factual consistency between textual information and visual evidence is critical for trustworthy MLLMs.

To alleviate object hallucination, prior methods have improved visual-text alignment through instruction-response tuning (Wang et al., 2024;

* Corresponding Author

Zhang et al., 2024a) or reinforcement learning (RL) (Yu et al., 2024; Sun et al., 2023). However, they often require extra training and datasets, increasing computational cost and complexity (Liu et al., 2024b; Bai et al., 2024; Zhao et al., 2024). Inference-stage methods provide lightweight alternatives. Contrastive decoding (Huang et al., 2024; Huo et al., 2024; Wang et al., 2024; Leng et al., 2024) mitigates hallucination by contrasting outputs from the original image and a perturbed variant, but it doubles inference time and may affect coherence. Visual attention modulation (Kang et al., 2025; Yin et al., 2025; Tu et al., 2025; Zhang et al., 2024b) and visual feature retracing (Zou et al., 2024) instead strengthen reliance on visual information. However, these methods **overlook a critical issue**: *simply increasing visual dependency may introduce unintended side effects in the model’s output*.

To address this gap, we investigate how visual attention affects object hallucination in open-ended MLLM generation. We identify a critical, yet **previously underexplored** trade-off: *while stronger visual dependency enhances an MLLM’s ability to recognize true objects, it simultaneously increases the likelihood of generating hallucinated ones*. As illustrated in **Figure 1**, compared to output under the lower visual attention, the model tends to append *new* hallucinated descriptions in the trailing part of the generated sentence under stronger visual attention. We term this phenomenon *additive hallucination*. In **Section 3**, we quantify this phenomenon using positioning analysis and show that *additive hallucinations* consistently increase as visual attention strengthens, suggesting that **excessive emphasis on visual features** can inadvertently introduce new forms of hallucination rather than eliminating them.

Motivated by this observation, we propose *Global-to-Local Visual Attention Correction* (GLVAC), a **training-free inference-time** framework that decouples robust perception from precise description through a global-to-local dual-path architecture. The global pathway produces a conservative initial response using moderated visual attention, thereby reducing *additive hallucinations*, while the local correction branch **selectively** increases visual attention when uncertainty arises to correct hallucinations induced by language priors. A gating mechanism applies correction only when the main path is uncertain and the correction branch provides a

more confident prediction, enabling safe recovery of suppressed details without amplifying hallucinations caused by excessive visual attention. In summary, our main **contributions** are as follows:

- We characterize a fundamental **trade-off** between grounding completeness and visual factuality in open-ended generation of MLLMs, and identify *additive hallucination*, a new phenomenon arising under excessive visual attention.
- We introduce **GLVAC**, a *plug-and-play, training-free inference-time* method that dynamically regulates visual attention to balance robust global perception and local detail correction.
- We evaluate **GLVAC** across multiple MLLMs and benchmarks, showing that balancing global and local visual attention reduces both *language-prior hallucinations* and *additive hallucinations*.

2 Related Work

The emergence of MLLMs extends the language capabilities of LLMs to visual understanding and reasoning (Li et al., 2023a; Liu et al., 2024c; Alayrac et al., 2022). Typically, an MLLM integrates a pre-trained vision encoder, an LLM, and a learnable connector to align visual and textual representations; for example, LLaVA maps CLIP visual features into the LLM embedding space through a linear projector (Yin et al., 2024; Liu et al., 2024c; Radford et al., 2021). Despite their progress, MLLMs still suffer from **hallucination** (Jiang et al., 2024), where generated text is inconsistent with the input image. In particular, **object hallucination** refers to describing plausible but absent objects, which undermines reliability and trustworthiness, especially in safety-critical scenarios (Gunjal et al., 2024; Li et al., 2023b; Yoon et al., 2025). Mitigating such hallucinations is therefore essential for robust real-world deployment.

Existing hallucination-mitigation methods for MLLMs can be broadly divided into **training-based** and **inference-based** approaches (Bai et al., 2024; Wang et al., 2024; Huang et al., 2024). Training-based methods improve visual grounding through hallucination-sensitive instruction tuning, reinforcement learning with human feedback, or

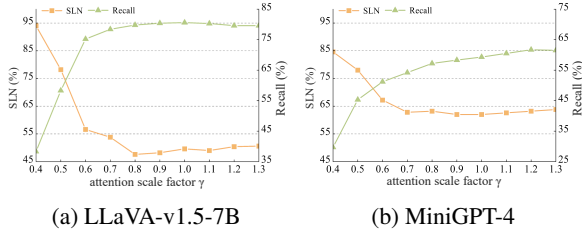


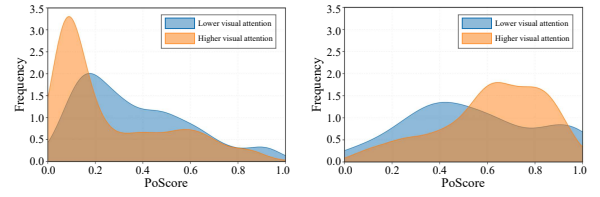
Figure 2: Effect of attention scale factor γ on Sentence-Level Non-Hallucination Rate (SLN) and Object Recall across two MLLMs, LLaVA-v1.5-7B and MiniGPT-4 with Llama 2 Chat 7B. SLN measures the proportion of responses containing no hallucinated objects. On both models, increasing γ improves object recall but reduces SLN, revealing a consistent trade-off between grounding completeness and visual factual consistency.

contrastive learning, but usually require costly annotation and retraining (Zhang et al., 2024a; Yu et al., 2024; Jiang et al., 2024). Inference-based methods avoid additional training by modifying decoding behavior, including **contrastive decoding**, **attention modulation**, and visual-token reinjection as in MEMVR (Huang et al., 2024; Kang et al., 2025; Zou et al., 2024). Although efficient, these methods mainly emphasize strengthening visual focus while overlooking its possible side effects. In this work, we further examine this issue and propose a simple yet effective approach for more comprehensive hallucination mitigation. A component-level comparison of GLVAC with VCD, VAR, and MEMVR is provided in **Appendix A**.

3 Observation: Trade-off Effects of Visual Attention in MLLMs

It is widely assumed that hallucinations in MLLMs primarily arise from an over-reliance on linguistic priors. To mitigate this, a growing body of works enhances visual grounding by amplifying cross-modal attention or retrieving salient visual features, under the belief that stronger visual dependence leads to more reliable generation.

To better understand **how** visual attention affects the model’s behavior, we vary the attention strength using a scaling factor γ , and larger values correspond to stronger visual focus on the image. As illustrated in **Figure 2**, both LLaVA-v1.5-7B and MiniGPT-4 show that when γ increases from low to moderate levels, object recall, defined as the proportion of ground-truth objects correctly mentioned, rises sharply, indicating more effective use of visual evidence. In contrast, the *Sentence-Level Non-Hallucination Rate* (SLN), defined as the pro-



(a) Ground-truth Object Positioning Score (b) Hallucinated Object Positioning Score

Figure 3: Positioning Score distributions for ground-truth objects (left) and hallucinated objects (right) under lower and higher levels of visual attention. Lower-attention and higher-attention settings correspond to smaller and larger values of the attention scaling factor γ , respectively.

portion of responses that contain no hallucinated objects, drops rapidly over the same range. Beyond this initial regime, recall largely plateaus, while SLN remains at a low level with only minor fluctuations as γ continues to increase. The inverse trend in the region where the model is most sensitive to attention scaling reveals a fundamental **trade-off between grounding completeness and factual consistency**: stronger visual attention improves the model’s use of visual evidence, yet it also makes the model more prone to hallucinating nonexistent objects.

To analyze **where** the increasing hallucinations emerge, we adopt the *Positioning Score* introduced in previous work (Zhou et al., 2023) and use it to study object placement in our setting. For a response s with length N_s , the score is

$$\text{PoScore}_{s,i} = \frac{\text{Index}(o_{s,i})}{N_s}. \quad (1)$$

Here, $o_{s,i}$ is the i^{th} mentioned object in response s . The term *Index* refers to the position of this object in the response measured in words. A higher score indicates a later occurrence in the response. For each object, we record only the first position at which it appears.

We examine the Positioning Score distributions under two levels of visual attention, as shown in **Figure 3**. Ground-truth objects appear early in the responses, and they occur more frequently under stronger visual attention. Hallucinated objects exhibit a different pattern. Under stronger attention, their Positioning Scores shift toward later parts of the responses. These hallucinations are added near the end of the responses instead of modifying the earlier correct content. We refer to them as *additive hallucinations*. This pattern shows that stronger vi-

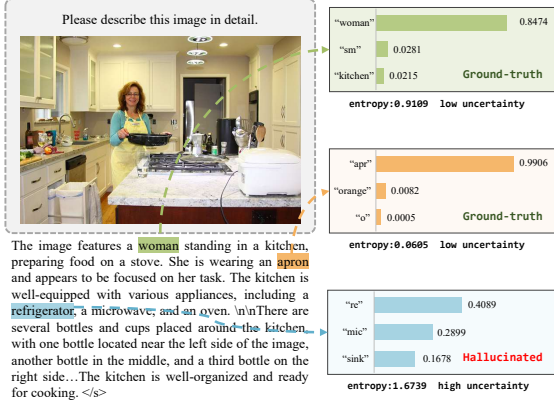


Figure 4: Token-level entropy patterns for ground-truth vs. hallucinated objects. Ground-truth tokens (e.g., “woman”, “apron”) show low entropy and high confidence. Hallucinated tokens (e.g., “refrigerator”) appear at high-entropy steps where the model shows uncertainty. This pattern indicates that hallucinations tend to arise in locally uncertain regions during generation.

sual attention preserves early correct descriptions but encourages additional, nonexistent objects in the later stages of generation. A controlled analysis under matched word and object-mention budgets is provided in **Appendix E.1**.

To further investigate **when** hallucinated objects emerge during generation, we analyze the token-level entropy of the model’s predictive distribution. As shown in **Figure 4**, entropy reflects the model’s uncertainty at each decoding step. We observe that hallucinated objects tend to occur at positions where the model exhibits high predictive entropy, suggesting that local uncertainty correlates with increased hallucination risk. This motivates using entropy as a lightweight signal to identify potentially unreliable decoding steps for selective correction, rather than applying correction uniformly.

These observations indicate that neither strong nor weak visual attention alone is sufficient for reliable generation. An adaptive, uncertainty-aware mechanism is therefore needed, which we introduce in the next section.

4 Global-to-Local Visual Attention Correction

Building on the insights presented in **Section 3**, we introduce *Global-to-Local Visual Attention Correction* (GLVAC), a training-free inference framework that adaptively regulates visual attention for more reliable multimodal reasoning. As illustrated in **Figure 5**, GLVAC decomposes visual reasoning

Algorithm 1 Global-to-Local Visual Attention Correction

Require: text-to-image submatrix Z_{img} of the pre-softmax self-attention logits, scale factors $\gamma_1 < \gamma_2$, entropy threshold δ

```

1: for each decoding step  $t$  do
2:   Global path (conservative visual attention)
3:    $Z'_{\text{img}} \leftarrow \gamma_1 Z_{\text{img}}$ 
4:   compute  $p_t^{(1)} = \text{Softmax}(\text{Logit}_1)$ 
5:    $H_t^{(1)} \leftarrow -\sum_{v \in V} p_t^{(1)}(v) \log p_t^{(1)}(v)$ 
6:   Uncertainty-triggered local correction
7:   if  $H_t^{(1)} > \delta$  then
8:     Local path (enhanced visual attention)
9:      $Z''_{\text{img}} \leftarrow \gamma_2 Z_{\text{img}}$ 
10:    compute  $p_t^{(2)} = \text{Softmax}(\text{Logit}_2)$ 
11:    Confidence gating
12:    if  $\max(p_t^{(2)}) > \max(p_t^{(1)})$  then
13:       $p_t^{\text{final}} \leftarrow p_t^{(2)}$   $\triangleright$  apply local correction
14:    else
15:       $p_t^{\text{final}} \leftarrow p_t^{(1)}$   $\triangleright$  keep global prediction
16:    end if
17:  else
18:     $p_t^{\text{final}} \leftarrow p_t^{(1)}$   $\triangleright$  skip local correction
19:  end if
20:  obtain token  $y_t$  from  $p_t^{\text{final}}$ 
21:  update the decoder state with  $y_t$ 
22: end for

```

into two complementary paths and selectively balances them based on model confidence. This design enables the model to dynamically trade off grounding completeness and accuracy, effectively mitigating hallucinations caused by both insufficient and excessive visual dependence.

4.1 Global Path: Conservative Visual Reasoning

To prevent hallucinations induced by excessive visual attention, GLVAC establishes a **global main path** that performs conservative visual reasoning as the default generation route.

Given a text token representation as the query Q and visual patch features as keys/values (K_v, V_v) , the decoder computes the pre-softmax self-attention logits

$$Z_{\text{img}} = \frac{QK_v^\top}{\sqrt{d}}, \quad (2)$$

where d denotes the dimensionality of the query and key vectors. The decoder then obtains the visual context as

$$\text{Attn}(Q, K_v, V_v) = \text{softmax}(Z_{\text{img}}) V_v. \quad (3)$$

For clarity, we omit the multi-head notation and matrix form. This mechanism significantly influences how much each next-token prediction relies on visual evidence versus language information.

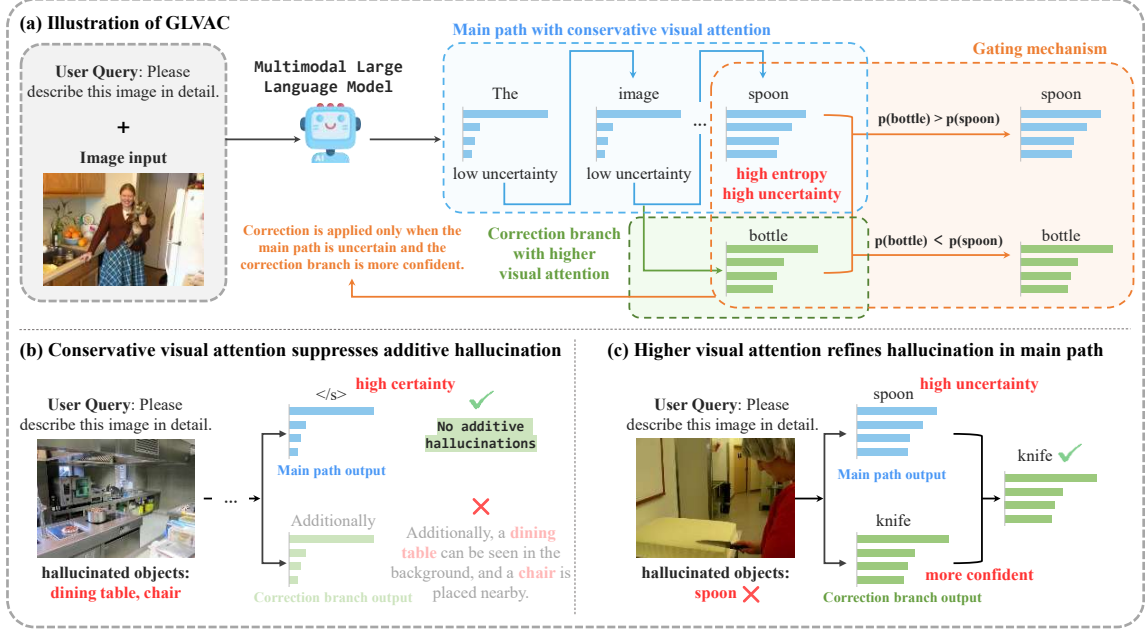


Figure 5: Illustration of the *Global-to-Local Visual Attention Correction* (GLVAC) framework. (a) The dual-path architecture: the main path uses conservative visual attention to avoid *additive hallucination*, while the correction path uses stronger visual attention to provide alternative candidates at uncertain decoding steps. A gating mechanism selects between the two paths based on entropy and confidence. (b) Conservative attention suppresses *additive hallucination* that may appear under stronger visual attention (e.g., “dining table”, “chair”). (c) Higher attention in the correction path helps fix hallucinated objects from the main path when it is more confident.

Therefore, we modulate visual reliance by scaling the text-to-image submatrix of the self-attention logits before the softmax.

Let $\mathbf{Z}_{\text{img}} \in \mathbb{R}^{L_t \times L_v}$ denote the pre-softmax attention score matrix from text tokens to image patches, where L_t and L_v are the numbers of text tokens and visual patches, respectively. Each element $\mathbf{Z}_{\text{img}}(t, v)$ indicates how strongly the t -th text token attends to the v -th visual patch, and the t -th row $\mathbf{Z}_{\text{img}}(t, :)$ thus represents the pre-softmax attention scores of the t -th text token over all visual regions.

To modulate the model’s reliance on visual signals without retraining, we scale the pre-softmax attention score matrix by an attention scale factor γ_1 :

$$\mathbf{Z}'_{\text{img}} = \gamma_1 \cdot \mathbf{Z}_{\text{img}}. \quad (4)$$

The scaled self-attention logits $\mathbf{Z}'_{\text{img}}$ are used to compute the token logits Logit_1 , and the resulting probability distribution is:

$$p^{(1)} = \text{Softmax}(\text{Logit}_1). \quad (5)$$

Here, γ_1 controls how much the model attends to image features during decoding: If γ_1 is too small, the model under-utilizes visual information

and risks generating language-prior-driven hallucinations; if γ_1 is too large, the model over-focuses visual input, easily introducing *additive hallucinations*.

Therefore, γ_1 is chosen to suppress additive hallucinations while preserving sufficient global perceptual capacity for downstream correction. This ensures that the global path produces *stable and correctable* predictions, serving as the foundation for subsequent local refinement.

4.2 Local Correction Path: Adaptive Visual Enhancement

After using the global path to mitigate additive hallucination, GLVAC introduces a **local correction branch** that applies stronger visual attention scaling to recover details that may be hallucinated in the conservative path. Using a higher factor $\gamma_2 > \gamma_1$, the same text-to-image self-attention logits are rescaled as:

$$\mathbf{Z}''_{\text{img}} = \gamma_2 \cdot \mathbf{Z}_{\text{img}}, \quad (6)$$

producing a more visually grounded distribution:

$$p^{(2)} = \text{Softmax}(\text{Logit}_2). \quad (7)$$

This correction branch is not used for direct generation. Instead, it acts as an auxiliary module

that provides refined predictions when the global path becomes uncertain. The higher visual scaling γ_2 amplifies the contribution of image features, increasing reliance on visual input. However, since excessive attention can also introduce hallucinations, its activation must be carefully controlled through the uncertainty-confidence gating mechanism described next.

Table 1: Performance evaluation on CHAIR. GLVAC achieves the lowest CHAIRs and CHAIRi scores across all models, demonstrating effective hallucination suppression, while also improving object recall compared with regular decoding. The best results are **bolded**. **Green** denotes improvement, and **Red** means degradation.

Model	Method	CHAIRs ↓	CHAIRi ↓	Recall ↑
LLaVA-v1.5-7B	regular	50.4 ↑0.0	14.2 ↑0.0	80.6 ↑0.0
	VCD	51.2 ↑0.8	15.1 ↑0.9	80.2 ↓0.4
	VAR	52.2 ↑1.8	14.9 ↑0.7	82.2 ↑1.6
	MEMVR	52.6 ↑2.2	14.9 ↑0.7	81.2 ↑0.6
	GLVAC	46.2 ↓4.2	13.9 ↓0.3	81.2 ↑0.6
LLaVA-v1.5-13B	regular	48.2 ↑0.0	13.2 ↑0.0	79.4 ↑0.0
	VCD	49.0 ↑0.8	13.7 ↑0.5	80.4 ↑1.0
	VAR	49.6 ↑1.4	13.5 ↑0.3	81.3 ↑1.9
	MEMVR	46.2 ↓2.0	12.5 ↓0.7	80.8 ↑1.4
	GLVAC	44.4 ↓3.8	12.2 ↓1.0	80.9 ↑1.5
MiniGPT-4	regular	37.6 ↑0.0	14.4 ↑0.0	59.4 ↑0.0
	VCD	37.8 ↑0.2	14.9 ↑0.5	60.5 ↑1.1
	VAR	38.0 ↑0.4	14.5 ↑0.1	59.3 ↓0.1
	MEMVR	39.0 ↑1.4	15.0 ↑0.6	59.8 ↑0.4
	GLVAC	36.6 ↓1.0	11.7 ↓2.7	63.5 ↑4.1

Table 2: Performance evaluation on AMBER generative task. GLVAC achieves the lowest CHAIR and Cog scores across all models, demonstrating effective hallucination suppression and better cognitive alignment. The best results are **bolded**. **Green** denotes improvement, and **Red** means degradation.

Model	Method	Cover ↑	CHAIR ↓	Hal ↓	Cog ↓
LLaVA-v1.5-7B	regular	51.7 ↑0.0	7.7 ↑0.0	35.8 ↑0.0	4.2 ↑0.0
	VCD	51.8 ↑0.1	8.8 ↑1.1	40.2 ↑4.4	4.0 ↓0.2
	VAR	52.8 ↑1.1	8.1 ↑0.4	38.1 ↑2.3	4.3 ↑0.1
	MEMVR	52.2 ↑0.5	7.8 ↑0.1	36.9 ↑1.1	4.3 ↑0.1
	GLVAC	51.3 ↓0.4	7.3 ↓0.4	33.4 ↓2.4	3.9 ↓0.3
LLaVA-v1.5-13B	regular	51.9 ↑0.0	6.8 ↑0.0	31.1 ↑0.0	3.3 ↑0.0
	VCD	52.0 ↑0.1	8.8 ↑2.0	40.3 ↑9.2	4.1 ↑0.8
	VAR	52.0 ↑0.1	7.4 ↑0.6	34.8 ↑3.7	3.7 ↑0.4
	MEMVR	52.2 ↑0.3	6.9 ↑0.1	31.4 ↑0.3	3.5 ↑0.2
	GLVAC	53.2 ↑1.3	6.4 ↓0.4	31.3 ↑0.2	3.0 ↓0.3
MiniGPT-4	regular	60.8 ↑0.0	19.2 ↑0.0	72.6 ↑0.0	12.1 ↑0.0
	VCD	60.5 ↓0.3	20.0 ↑0.8	72.0 ↓0.6	11.9 ↓0.2
	VAR	60.9 ↑0.1	19.0 ↓0.2	72.7 ↑0.1	12.1 ↑0.0
	MEMVR	61.2 ↑0.4	18.6 ↓0.6	71.3 ↓1.3	11.9 ↓0.2
	GLVAC	63.9 ↑3.1	15.1 ↓4.1	66.6 ↓6.0	10.8 ↓1.3

4.3 Uncertainty-Confidence Gating Mechanism

A central challenge is determining *when* the local correction path should intervene. GLVAC adopts

a dual-condition gating mechanism based on token-level uncertainty and relative confidence.

We use token-level predictive entropy as an uncertainty signal to identify potentially unreliable decoding positions. At each decoding step, GLVAC first obtains the main-path distribution and computes its predictive entropy:

$$H(p_t^{(1)}) = - \sum_{v \in \mathcal{V}} p_t^{(1)}(v) \log p_t^{(1)}(v) \quad (8)$$

If $H(p_t^{(1)}) \leq \delta$, the local correction path is skipped and the global prediction is directly retained. Only when $H(p_t^{(1)}) > \delta$ do we execute the local correction path to obtain $p_t^{(2)}$. To avoid unnecessary or harmful corrections, the local prediction replaces the global prediction only when it provides a higher confidence:

$$\max(p_t^{(2)}) > \max(p_t^{(1)}) \quad (9)$$

The final token distribution is thus selected as:

$$p_t^{\text{final}} = \begin{cases} p_t^{(2)}, & \text{if } H(p_t^{(1)}) > \delta \wedge \\ & \max(p_t^{(2)}) > \max(p_t^{(1)}), \\ p_t^{(1)}, & \text{otherwise.} \end{cases} \quad (10)$$

This gating ensures that the correction occurs only at high-uncertainty positions and only when a more reliable alternative is available. As a result, GLVAC avoids over-attention-induced hallucinations, prevents unnecessary interventions, and provides adaptive token-level control with a controlled computational overhead. The complete inference procedure is summarized in Algorithm 1.

5 Experiment

5.1 Experimental Settings

Benchmarks and Metrics. We evaluate GLVAC on several widely recognized benchmarks to assess its effectiveness in mitigating hallucinations and preserving grounding in MLLMs: CHAIR (Rohrbach et al., 2018), AMBER (Wang et al., 2023a), and MM-Vet (Yu et al., 2023).

Backbones and Baselines. We evaluate GLVAC across three representative LVLMs: **LLaVA-v1.5-7B**, **LLaVA-v1.5-13B** (Liu et al., 2024c), and **MiniGPT-4 with Llama 2 Chat 7B** (Zhu et al., 2023). For comparison, we adopt three training-free baselines: **VCD** (Leng et al., 2024), **VAR** (Kang et al., 2025), and **MEMVR** (Zou et al., 2024).

Table 3: Performance evaluation on MM-Vet, where rec denotes recognition, know denotes knowledge, gen denotes language generation, and spat denotes spatial awareness. GLVAC consistently improves the overall performance of all three MLLMs, highlighting its effectiveness in advancing general-purpose multimodal understanding. The best results are **bolded**. **Green** denotes improvement, and **Red** means degradation.

Model	Method	rec $\uparrow$	ocr $\uparrow$	know $\uparrow$	gen $\uparrow$	spat $\uparrow$	math $\uparrow$	Total $\uparrow$
LLaVA-v1.5-7B	regular	36.7 $\uparrow$ 0.0	24.8 $\uparrow$ 0.0	19.3 $\uparrow$ 0.0	23.4 $\uparrow$ 0.0	27.1 $\uparrow$ 0.0	11.5 $\uparrow$ 0.0	31.9 $\uparrow$ 0.0
	VCD	34.6 $\downarrow$2.1	22.1 $\downarrow$2.7	18.6 $\downarrow$0.7	19.7 $\downarrow$3.7	27.5 $\uparrow$0.4	6.9 $\downarrow$4.6	30.1 $\downarrow$1.8
	VAR	38.9 $\uparrow$2.2	24.2 $\downarrow$0.6	22.0 $\uparrow$2.7	21.6 $\downarrow$1.8	28.7 $\uparrow$1.6	3.8 $\downarrow$7.7	34.0 $\uparrow$2.1
	MEMVR	36.9 $\uparrow$0.2	24.5 $\downarrow$0.3	18.9 $\downarrow$0.4	22.7 $\downarrow$0.7	25.9 $\downarrow$1.2	11.2 $\downarrow$0.3	32.2 $\uparrow$0.3
	GLVAC	37.7 $\uparrow$1.0	24.4 $\downarrow$0.4	21.5 $\uparrow$2.2	22.5 $\downarrow$0.9	29.5 $\uparrow$2.4	11.5 $\uparrow$0.0	33.3 $\uparrow$1.4
LLaVA-v1.5-13B	regular	40.1 $\uparrow$ 0.0	30.4 $\uparrow$ 0.0	25.1 $\uparrow$ 0.0	24.6 $\uparrow$ 0.0	33.3 $\uparrow$ 0.0	15.0 $\uparrow$ 0.0	36.9 $\uparrow$ 0.0
	VCD	38.5 $\downarrow$1.6	27.6 $\downarrow$2.8	25.4 $\uparrow$0.3	25.4 $\uparrow$0.8	33.9 $\uparrow$0.6	11.2 $\downarrow$3.8	35.0 $\downarrow$1.9
	VAR	41.1 $\uparrow$1.0	29.9 $\downarrow$0.5	24.9 $\downarrow$0.2	26.4 $\uparrow$1.8	34.7 $\uparrow$1.4	11.2 $\downarrow$3.8	37.0 $\uparrow$0.1
	MEMVR	40.8 $\uparrow$0.7	32.8 $\uparrow$2.4	25.1 $\uparrow$ 0.0	26.1 $\uparrow$1.5	36.0 $\uparrow$2.7	15.0 $\uparrow$ 0.0	37.9 $\uparrow$1.0
	GLVAC	42.0 $\uparrow$1.9	32.2 $\uparrow$1.8	25.6 $\uparrow$0.5	26.6 $\uparrow$2.0	37.3 $\uparrow$4.0	14.2 $\downarrow$0.8	38.6 $\uparrow$1.7
MiniGPT-4	regular	30.8 $\uparrow$ 0.0	20.2 $\uparrow$ 0.0	21.7 $\uparrow$ 0.0	20.1 $\uparrow$ 0.0	28.3 $\uparrow$ 0.0	9.2 $\uparrow$ 0.0	26.7 $\uparrow$ 0.0
	VCD	25.7 $\downarrow$5.1	23.8 $\uparrow$3.6	19.5 $\downarrow$2.2	18.2 $\downarrow$1.9	31.6 $\uparrow$3.3	12.3 $\uparrow$3.1	25.0 $\downarrow$1.7
	VAR	30.7 $\downarrow$0.1	19.6 $\downarrow$0.6	21.9 $\uparrow$0.2	19.7 $\downarrow$0.4	27.6 $\downarrow$0.7	9.2 $\uparrow$ 0.0	26.6 $\downarrow$0.1
	MEMVR	31.2 $\uparrow$0.4	21.7 $\uparrow$1.5	23.3 $\uparrow$1.6	21.2 $\uparrow$1.1	30.0 $\uparrow$1.7	8.5 $\downarrow$0.7	27.4 $\uparrow$0.7
	GLVAC	32.3 $\uparrow$1.5	21.8 $\uparrow$1.6	22.7 $\uparrow$1.0	22.1 $\uparrow$2.0	28.8 $\uparrow$0.5	11.9 $\uparrow$2.7	28.1 $\uparrow$1.4

Implementation Details. We use greedy decoding for all experiments and follow the original settings for VCD (Leng et al., 2024), VAR (Kang et al., 2025), and MEMVR (Zou et al., 2024). For MEMVR settings not covered by the original experiments, we grid-search the retracing ratio with a step size of 0.01. In GLVAC, we fix the entropy threshold at 1.5 for all tasks. More details are provided in **Appendix B**.

5.2 Main Results and Analysis

Results on CHAIR. Table 1 reports the results on the CHAIR benchmark. GLVAC consistently achieves the lowest CHAIRs and CHAIRi scores across all MLLMs. The improvement is most evident on the LLaVA series, where GLVAC reduces CHAIRs by 4.2% on LLaVA-v1.5-7B and 3.8% on LLaVA-v1.5-13B, respectively. These results demonstrate that GLVAC effectively mitigates object hallucination by dual-path visual attention modulation.

Results on AMBER. Table 2 presents the results on the AMBER generative benchmark. Compared to the baselines, GLVAC consistently achieves the lowest CHAIR and Cog scores. On LLaVA-v1.5-7B, GLVAC lowers Hal by 2.4% while keeping coverage comparable to the regular model. On LLaVA-v1.5-13B, it simultaneously improves coverage and further reduces CHAIR and Cog. The gains are most pronounced on MiniGPT-4, where GLVAC boosts coverage by 3.1% and

reduces Hal by 6.0%. These results show that GLVAC mitigates hallucinations without sacrificing object coverage, and can improve coverage on stronger backbones.

Results on MM-Vet. Table 3 presents the results on the MM-Vet benchmark. GLVAC improves the overall score for all three MLLMs. On LLaVA-v1.5-7B and LLaVA-v1.5-13B, it increases the overall score by 1.4 and 1.7 points, with clear gains in recognition, knowledge, and spatial reasoning. The largest improvement is observed on MiniGPT-4, where GLVAC raises the total score by 1.4 points and enhances all scores. These results indicate that the proposed method strengthens general multimodal understanding in addition to reducing hallucinations.

More experimental results and an inference-efficiency analysis are provided in **Appendices C and D**.

5.3 Ablation Study and Analysis

Effect of Dual-Path Design: Global-Only vs. GLVAC. To verify the necessity of the dual-path mechanism, we compare GLVAC with a Global-Only variant that applies a single global scaling factor γ without the correction branch. As shown in Table 4, Global-Only exhibits a clear trade-off: smaller γ reduces hallucinations but hurts recall, while larger γ improves recall at the cost of more hallucinations. No single γ yields consistently strong performance across metrics. In

contrast, GLVAC achieves the best overall balance by using a conservative global path while selectively invoking local correction when needed.

Table 4: Ablation study of correction branch using LLaVA-v1.5-7B on the CHAIR benchmark. The first group shows performance with varying γ values without the correction branch. The with-correction-branch variant (GLVAC) uses scaling parameters $\gamma_1 = 0.8$ and $\gamma_2 = 1.2$, which achieves the best performance in both CHAIRs and Recall. The best results are bolded.

Method	CHAIRs ↓	CHAIRi ↓	Recall ↑	
without correction branch	$\gamma = 0.7$	46.2	13.3	78.4
	$\gamma = 0.8$	52.4	14.8	79.8
	$\gamma = 0.9$	51.8	14.6	80.4
	$\gamma = 1.0$	50.4	14.2	80.6
	$\gamma = 1.1$	51.0	13.9	80.3
	$\gamma = 1.2$	49.6	13.7	79.6
	$\gamma = 1.3$	49.4	14.0	79.6
with correction branch	46.2	13.9	81.2	

Ablation on Gating Conditions. We further ablate the gating strategy with entropy-only, confidence-only, and dual-condition variants. As shown in **Table 5**, both single-signal variants improve over the baseline, but each has limitations: Entropy-only allows all corrections from the branch to replace the main path’s output whenever the correction branch is activated, potentially introducing hallucinations from the correction path while correcting the main path’s hallucinations. On the other hand, confidence-only, due to the absence of entropy constraints, causes frequent activation of the correction path, significantly increasing the inference time. Our dual-condition gating delivers the best effectiveness–efficiency trade-off, achieving the lowest CHAIRs and CHAIRi.

Table 5: Ablation study on gating-condition strategies using LLaVA-v1.5-13B on the CHAIR benchmark. Our method, GLVAC (dual-condition), achieves the best performance in both CHAIRs and CHAIRi, demonstrating the effectiveness of combining entropy and confidence for thresholding. The best results are bolded.

Method	CHAIRs ↓	CHAIRi ↓	Recall ↑
regular	50.0	13.7	79.8
entropy-only	47.6	13.4	80.1
confidence-only	47.2	12.4	81.2
dual-condition	44.4	12.2	80.9

Ablation on Entropy Threshold. **Table 6** shows the effect of the entropy threshold δ . A too small δ (e.g., 0.6) over-triggers correction, leading to higher CHAIRs due to *additive hallucina-*

tions; a large δ (e.g., 2.1) rarely activates correction, degrading performance toward the global-only baseline. Moderate δ (1.2–1.8) balances hallucination suppression and recall. Importantly, GLVAC outperforms the regular baseline across a wide range of δ (0.6–2.1), demonstrating that it is relatively robust and less sensitive to the choice of δ within this range. These results validate our uncertainty-guided gating: a moderate entropy threshold enables GLVAC to selectively apply local correction on truly uncertain tokens, rather than aggressively modifying confident predictions or completely turning off the correction branch.

Table 6: Ablation of entropy threshold values using LLaVA-v1.5-13B on the CHAIR benchmark. GLVAC outperforms the regular baseline across a wide range of entropy threshold values (0.6–2.1). The best results are bolded.

Threshold	CHAIRs ↓	CHAIRi ↓	Recall ↑
0.6	47.2	12.4	81.2
0.9	45.4	12.3	81.1
1.2	45.6	12.3	81.3
1.5	44.4	12.2	80.9
1.8	44.2	12.1	80.6
2.1	45.4	12.4	80.3

Additional analyses of the global-path and local-path attention scaling factors and the layer-selection strategy are provided in **Appendix E**.

6 Conclusion

In this work, we uncover a key trade-off in MLLMs: stronger visual attention improves visual grounding completeness but also increases the frequency of additive hallucinations. To address this issue, we propose GLVAC, a training-free dual-path inference framework. GLVAC employs a conservative global path to suppress additive hallucinations and a selectively activated local correction branch that applies stronger visual attention only at high-uncertainty decoding steps, guided by entropy and confidence thresholds. Extensive experiments across multiple MLLMs and benchmarks demonstrate that GLVAC effectively reduces object hallucinations while preserving or even improving grounding completeness.

Limitations

Although GLVAC is training-free and plug-and-play, it has two limitations. First, because the

framework introduces an additional local correction path at uncertain decoding steps, its inference time is inevitably higher than that of the original model. Although the gating mechanism controls this overhead by activating correction selectively, latency remains a concern for real-time applications. Second, GLVAC relies on several hyperparameters, including the global and local attention scaling factors and the entropy threshold. These parameters may vary across model architectures and tasks. Developing adaptive parameter selection strategies is left for future work.

References

- Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, and 1 others. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736.
- Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.
- Davide Bucciarelli, Nicholas Moratelli, Marcella Cornia, Lorenzo Baraldi, and Rita Cucchiara. 2024. Personalizing multimodal large language models for image captioning: an experimental analysis. In *European Conference on Computer Vision*, pages 351–368. Springer.
- Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, and 1 others. 2024. A survey on multimodal large language models for autonomous driving. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 958–979.
- Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2025. 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*.
- Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 18135–18143.
- MD Zakir Hossain, Ferdous Sohel, Mohd Fairuz Shiratuddin, and Hamid Laga. 2019. A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, 51(6):1–36.
- Minbin Huang, Yanxin Long, Xincheng Deng, Ruihang Chu, Jiangfeng Xiong, Xiaodan Liang, Hong Cheng, Qinglin Lu, and Wei Liu. 2025. Dialoggen: Multimodal interactive dialogue system with multi-turn text-image generation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 411–426.
- Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. 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.
- Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2024. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*.
- Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.
- Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. 2025. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*.
- Jiayi Kuang, Ying Shen, Jingyou Xie, Haohao Luo, Zhe Xu, Ronghao Li, Yinghui Li, Xianfeng Cheng, Xika Lin, and Yu Han. 2025. Natural language understanding and inference with mllm in visual question answering: A survey. *ACM Computing Surveys*, 57(8):1–36.
- Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. 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.
- Chunyu Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36:28541–28564.
- Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*.
- Chenxiao Liu, Zheyong Xie, Sirui Zhao, Jin Zhou, Tong Xu, Minglei Li, and Enhong Chen. 2024a.

- Speak from heart: an emotion-guided llm-based multimodal method for emotional dialogue generation. In *Proceedings of the 2024 International Conference on Multimedia Retrieval*, pages 533–542.
- Fenglin Liu, Tingting Zhu, Xian Wu, Bang Yang, Chenyu You, Chenyang Wang, Lei Lu, Zhangdaihong Liu, Yefeng Zheng, Xu Sun, and 1 others. 2023. A medical multimodal large language model for future pandemics. *NPJ Digital Medicine*, 6(1):226.
- Hanchao Liu, Wenyuan Xue, Yifei Chen, Dapeng Chen, Xiutian Zhao, Ke Wang, Liping Hou, Rongjun Li, and Wei Peng. 2024b. A survey on hallucination in large vision-language models. *arXiv preprint arXiv:2402.00253*.
- Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.
- Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmlR.
- Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*.
- Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2023. Aligning large multimodal models with factually augmented rlhf. *arXiv preprint arXiv:2309.14525*.
- Chongjun Tu, Peng Ye, Dongzhan Zhou, Lei Bai, Gang Yu, Tao Chen, and Wanli Ouyang. 2025. Attention reallocation: Towards zero-cost and controllable hallucination mitigation of mllms. *arXiv preprint arXiv:2503.08342*.
- Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and 1 others. 2023a. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.
- Lei Wang, Jiabang He, Shenshen Li, Ning Liu, and Ee-Peng Lim. 2024. Mitigating fine-grained hallucination by fine-tuning large vision-language models with caption rewrites. In *International Conference on Multimedia Modeling*, pages 32–45. Springer.
- Wenhai Wang, Jiangwei Xie, ChuanYang Hu, Haoming Zou, Jianan Fan, Wenwen Tong, Yang Wen, Silei Wu, Hanming Deng, Zhiqi Li, and 1 others. 2023b. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. *arXiv preprint arXiv:2312.09245*.
- Dexuan Xu, Yanyuan Chen, Jieyi Wang, Yue Huang, Hanpin Wang, Zhi Jin, Hongxing Wang, Weihua Yue, Jing He, Hang Li, and 1 others. 2024. Mlevlm: Improve multi-level progressive capabilities based on multimodal large language model for medical visual question answering. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4977–4997.
- Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, and 1 others. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.
- Hao Yin, Guangzong Si, and Zilei Wang. 2025. Clear-sight: 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.
- Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. 2024. A survey on multimodal large language models. *National Science Review*, 11(12):nwae403.
- Dokyoon Yoon, Youngsook Song, and Woomyoung Park. 2025. Stop learning it all to mitigate visual hallucination, focus on the hallucination target. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 4200–4208.
- Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and 1 others. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.
- Weihaoyu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.
- Jinrui Zhang, Teng Wang, Haigang Zhang, Ping Lu, and Feng Zheng. 2024a. Reflective instruction tuning: Mitigating hallucinations in large vision-language models. In *European Conference on Computer Vision*, pages 196–213. Springer.
- Xiaofeng Zhang, Yihao Quan, Chaochen Gu, Chen Shen, Xiaosong Yuan, Shaotian Yan, Hao Cheng, Kaijie Wu, and Jieping Ye. 2024b. Seeing clearly by layer two: Enhancing attention heads to alleviate hallucination in vlms. *arXiv preprint arXiv:2411.09968*.
- Linxi Zhao, Yihe Deng, Weitong Zhang, and Quanquan Gu. 2024. Mitigating object hallucination in large

vision-language models via image-grounded guidance. *arXiv preprint arXiv:2402.08680*.

Juexiao Zhou, Xiaonan He, Liyuan Sun, Jiannan Xu, Xiuying Chen, Yuetan Chu, Longxi Zhou, Xingyu Liao, Bin Zhang, Shawn Afvari, and 1 others. 2024. Pre-trained multimodal large language model enhances dermatological diagnosis using skinpt-4. *Nature Communications*, 15(1):5649.

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

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

Xin Zou, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Ken-ting Zheng, Sirui Huang, Junkai Chen, Peijie Jiang, Jia Liu, Chang Tang, and 1 others. 2024. Look twice before you answer: Memory-space visual retracing for hallucination mitigation in multimodal large language models. *arXiv preprint arXiv:2410.03577*.

Appendix

A Component-Level Comparison with Inference-Time Methods

VCD (Leng et al., 2024), VAR (Kang et al., 2025), MEMVR (Zou et al., 2024), and GLVAC all mitigate hallucinations at inference time, but they intervene at different model components and embody different assumptions about visual reliance. VCD contrasts output logits obtained from the original and distorted images at every decoding step. VAR redistributes post-softmax attention mass from attention-sink tokens to non-sink visual tokens in selected heads. MEMVR uses intermediate-layer entropy to trigger visual-feature reinjection into the feed-forward network representations within a single prediction trajectory. These methods primarily contrast, redistribute, or restore visual evidence. In comparison, GLVAC explicitly considers the complementary risk introduced by excessive visual reliance. It establishes a global main path for conservative visual reasoning by scaling the text-to-image submatrix Z_{img} of the pre-softmax self-attention logits with γ_1 , and executes the local correction path with γ_2 only at high-uncertainty decoding steps. The local prediction

replaces the global prediction only when it provides a higher confidence.

B Experimental Setup Details

B.1 Datasets and Metrics

We evaluate our method on three widely used benchmarks for assessing multimodal hallucination and visual grounding quality: CHAIR (Rohrbach et al., 2018), AMBER (Wang et al., 2023a), and MM-Vet (Yu et al., 2023). All three benchmarks are English-language datasets. These benchmarks cover complementary aspects of hallucination, ranging from object-level inconsistencies to attribute-level reliability and holistic reasoning performance. Below, we describe each dataset and its evaluation metrics in detail.

CHAIR (Rohrbach et al., 2018). CHAIR evaluates object hallucination in image captioning. It focuses on whether objects mentioned in generated captions are grounded in the corresponding image. Following prior practice (Zou et al., 2024), we randomly sample 500 images from the full dataset for evaluation and query the MLLM with the prompt “Please describe this image in detail.”. CHAIR defines two hallucination metrics:

- **CHAIRi:** Instance-level hallucination rate, measuring the proportion of hallucinated object mentions:

$$\text{CHAIRi} = \frac{N_{\text{hall}}^{\text{obj}}}{N_{\text{all}}^{\text{obj}}}. \quad (11)$$

- **CHAIRs:** Sentence-level hallucination rate, measuring the proportion of captions containing at least one hallucinated object:

$$\text{CHAIRs} = \frac{N_{\text{hall}}^{\text{cap}}}{N_{\text{all}}^{\text{cap}}}. \quad (12)$$

In addition to hallucination rates, we also report the **object recall** of the generated captions to measure how well the generated captions cover the objects present in the image. **Recall** is defined as:

$$\text{Recall} = \frac{|\{\text{correctly mentioned ground-truth objects}\}|}{|\{\text{all ground-truth objects}\}|}. \quad (13)$$

AMBER (Wang et al., 2023a). AMBER consists of two components: generative task (AMBER-Gen) and discriminative task (AMBER-Disc). Our

Table 7: Component-level comparison of representative training-free inference-time methods. “Conditional” means that the second path is executed only when the entropy condition is satisfied.

Method	Modified component	Uncertainty signal	Paths	Explicitly reduces visual reliance
VCD	Image input and output logits	No	Two, every step	No
VAR	Post-softmax attention weights	No	One	No
MEMVR	FFN representations via visual-feature reinjection	Intermediate-layer entropy	One	No
GLVAC	Pre-softmax text-to-image self-attention logits; output-distribution gating	Output entropy	Two, conditional	Yes

experiments use only AMBER-Gen, which contains 1,004 samples. Following the original setup, we use the prompt “Describe this image.” for all AMBER-Gen evaluations. AMBER-Gen reports four sub-metrics:

- **Cover**: the proportion of annotated objects mentioned in the response.
- **CHAIR**: the proportion of object mentions in the response that do not appear in the annotated object list.
- **Hal**: the proportion of responses that contain hallucinations.
- **Cog**: the proportion of object mentions in the response that belong to the predefined hallucination target set H_{obj} , used to examine whether the hallucination tendencies of MLLMs align with patterns commonly observed in human cognition.

MM-Vet (Yu et al., 2023). MM-Vet evaluates multimodal models on 218 complicated tasks that require the integration of multiple core vision-language capabilities. It defines six fundamental capabilities: *recognition*, *OCR*, *knowledge*, *language generation*, *spatial awareness*, and *math*, and constructs 16 integrated task types by combining these capabilities. Following the original setting, we adopt the default MM-Vet evaluation prompt, and model outputs are scored by an LLM-based evaluator conditioned on the question, ground truth, and prediction. Each sample is assigned a score between 0 and 1, and the final benchmark score is computed as the average over all samples. In addition to the overall score, MM-Vet also reports capability-wise results by averaging over samples associated with a specific capability or capability combination. For all MM-Vet evaluations, we use the official grading tool provided by the

benchmark authors, which implements a GPT-4.1-based evaluator following the protocol described in the original paper.

B.2 Implementation Details

We adopt greedy decoding for all experiments across all models. The reported results on all three benchmarks are obtained from a single run. For **VCD**, which contrasts outputs from the original image and a visually distorted image during decoding, we set `do_sample=False`, the noise step to 500, the plausibility constraint hyperparameter λ to 0.1, and the contrastive emphasis coefficient α to 1, following the default parameter settings from the original code and literature (Leng et al., 2024). For the **VAR** baseline, which identifies visual sink tokens and redistributes their attention to enhance image focus, we strictly follow the original implementation and set the threshold for selecting sink tokens to $\tau = 20$ and the portion of the attention weights from the sink tokens into the attention budget to $p = 0.6$ for all experiments. The task-dependent threshold ρ for selecting image-centric heads is set to $\rho = 0.8$ for MM-Vet and $\rho = 0.5$ for CHAIR and AMBER. Following the original VAR design, we do not modify the attention heads in the last layer. For the **MEMVR** baseline, which dynamically re-injects visual features during generation, we set `entropy-threshold` to 0.75 and apply retracing from layer 5 to layer 16 for all experiments. Regarding the retracing ratio, besides the default values provided in the official implementation, we additionally perform a grid search in $[0.01, 0.4]$ with a step size of 0.01, and report the best-performing value for each setting. GLVAC introduces two scalar hyperparameters, γ_1 and γ_2 , which scale the visual attention strength in the global and local paths, respectively. We explore (γ_1, γ_2) over a small discrete set of scaling values within a restricted range, un-

Table 8: Hyperparameter settings of (γ_1, γ_2) for different backbones and benchmarks.

Benchmark	Model	γ_1	γ_2
CHAIR	LLaVA-v1.5-7B	0.8	1.2
	LLaVA-v1.5-13B	0.7	1.1
	MiniGPT-4	1.7	2.2
AMBER	LLaVA-v1.5-7B	0.7	1.4
	LLaVA-v1.5-13B	0.6	1.3
	MiniGPT-4	1.7	2.2
MM-Vet	LLaVA-v1.5-7B	1.2	1.5
	LLaVA-v1.5-13B	1.1	1.5
	MiniGPT-4	1.0	1.2

der the constraint $\gamma_1 < \gamma_2$. The selected values for each backbone-benchmark pair are reported in **Table 8**. A sensitivity analysis on LLaVA-v1.5-13B under the CHAIR benchmark shows that GLVAC remains effective over a broad range of γ_1 and γ_2 values in our separate sweeps, suggesting that the method is not overly sensitive to these hyperparameters within reasonable ranges. For all experiments, we fix the entropy threshold used by the gating mechanism to $\delta = 1.5$. In addition, based on our layer-wise attention inspection, we apply GLVAC starting from transformer layer 2 and leave layers 0–1 untouched. All experiments of LLaVA-v1.5-7B are conducted on an NVIDIA RTX 4090 GPU, while all experiments of LLaVA-v1.5-13B and MiniGPT-4 (Llama 2 Chat 7B backbone) are conducted on an NVIDIA H200 GPU.

C Additional Experiments

To further understand how visual attention strength influences different diagnostic dimensions, we evaluate LLaVA-v1.5-7B on the hallucination subset of MME benchmark (Fu et al., 2025) while varying the visual-attention scaling coefficient γ . MME is a comprehensive evaluation benchmark covering 14 perception and cognition subtasks, where each sample uses a unified “yes/no” instruction format to enable objective measurement of fine-grained visual grounding abilities (e.g., existence, count, position, color). **Table 9** reports the performance under a lower-attention setting ($\gamma = 0.95$) and a higher-attention setting ($\gamma = 1.05$) on visual features. Interestingly, we observe that reducing visual attention slightly below the default level leads to consistently better scores in existence, count, and color recognition, while in-

Table 9: MME performance under different attention scaling strengths.

	visual attention	existence $\uparrow$	count $\uparrow$	position $\uparrow$	color $\uparrow$
regular		190.0 $\uparrow 0.0$	160.0 $\uparrow 0.0$	138.3 $\uparrow 0.0$	165.0 $\uparrow 0.0$
lower ($\gamma = 0.95$)		195.0 $\uparrow 5.0$	163.0 $\uparrow 3.0$	143.0 $\uparrow 4.7$	175.0 $\uparrow 10.0$
higher ($\gamma = 1.05$)		185.0 $\downarrow 5.0$	155.0 $\downarrow 5.0$	145.00 $\uparrow 6.7$	155.0 $\downarrow 10.0$

creasing attention improves only the position subtask. This pattern indicates that different MME dimensions respond differently to visual grounding strength: excessive attention amplification tends to introduce spurious visual cues that harm existence and count recognition, whereas moderate attention provides a more stable semantic signal. These findings further support the motivation of GLVAC, using a conservative global path and a selectively strengthened local path to balance multiple evaluation dimensions.

D Inference Efficiency

To quantify the computational overhead introduced by GLVAC, we conduct a controlled comparison with regular decoding, VCD, VAR, and MEMVR on LLaVA-v1.5-13B. All methods are evaluated under the same hardware, greedy-decoding configuration, implementation environment, and 80-token generation budget on an NVIDIA H200 GPU. We report wall-clock latency, peak GPU memory, and floating-point operations per sample. As shown in **Table 10**, GLVAC incurs $1.63\times$ the latency of regular decoding. It is more costly than VAR ($1.08\times$), but substantially faster than VCD ($2.04\times$) and MEMVR ($2.93\times$).

GLVAC requires 19.83 TFLOPs per sample, similar to regular decoding at 18.85 and VAR at 19.32. In contrast, VCD and MEMVR require 37.69 and 28.08 TFLOPs, respectively. Its peak memory usage remains effectively identical to regular decoding. These results show that GLVAC is training-free but not computation-free: its uncertainty-triggered local path introduces measurable latency, while avoiding the substantially larger computational cost of VCD and MEMVR.

E Additional Analysis

E.1 Controlled Analysis of Additive Hallucination

Figure 3 shows hallucinated objects shifting toward the response tail as visual attention increases. To test whether longer outputs or additional object mentions explain this pattern, we perform two

Table 10: Inference-efficiency comparison on LLaVA-v1.5-13B under the same 80-token generation budget. Latency is reported in milliseconds per sample, with ratios relative to regular decoding in parentheses. Peak memory is measured in MB, and computation is reported in TFLOPs per sample.

Method	Latency (ms)	Peak memory (MB)	TFLOPs
Regular	1,687 (1.00 $\times$)	26,135	18.85
VCD	3,435 (2.04 $\times$)	26,678	37.69
VAR	1,817 (1.08 $\times$)	26,130	19.32
MEMVR	4,941 (2.93 $\times$)	26,296	28.08
GLVAC	2,743 (1.63$\times$)	26,135	19.83

Table 11: CHAIRs (% , lower is better) for LLaVA-v1.5-13B on CHAIR-500 after controlling the retained response length or number of object mentions. “Original” denotes the untruncated outputs.

Controlled budget	$\gamma = 0.6$	$\gamma = 0.7$	$\gamma = 0.9$	$\gamma = 1.1$	$\gamma = 1.3$
Original	40.8	44.6	45.4	47.6	50.8
50 words	18.2	18.2	20.6	19.4	22.6
70 words	29.6	32.6	34.6	37.6	39.4
90 words	38.6	42.0	43.0	44.8	47.4
4 objects	23.6	22.4	24.2	25.6	29.4
6 objects	29.8	32.0	34.0	36.6	39.2
8 objects	36.4	39.8	40.0	42.0	45.4

post-hoc controls on 500 CHAIR images generated by LLaVA-v1.5-13B. For each γ , we truncate responses to word-prefix budgets (50, 70, or 90 words) or object-mention budgets (4, 6, or 8 CHAIR-recognized object mentions), then recompute CHAIRs. Table 11 reports CHAIRs.

CHAIRs increases overall with γ under both controls. The strictest 50-word and four-object settings contain non-monotonic fluctuations, so we do not claim pointwise monotonicity. Nevertheless, increasing γ from 0.6 to 1.3 raises CHAIRs by 4.4, 9.8, and 8.8 points at 50, 70, and 90 words, respectively, and by 5.8, 9.4, and 9.0 points at four, six, and eight object mentions. Untruncated outputs show the same trend (a 10.0-point increase). Thus, response length and the number of retained object mentions do not by themselves explain the higher hallucination rate under stronger visual attention. Together with the positional shift in Figure 3, these controls strengthen the evidence for the additive-hallucination pattern.

E.2 Entropy Effectiveness and Local-Path Usage

Mention-level entropy analysis. We quantify whether predictive entropy distinguishes hallucinated from grounded object mentions using 100 CHAIR samples generated by LLaVA-v1.5-13B.

The analysis contains 831 object mentions: 122 hallucinated and 709 grounded. For multi-token mentions, the mention-level score is the maximum token entropy over the aligned span. Table 12 shows that this score achieves an AUROC of 0.735. Hallucinated mentions have a higher mean entropy than grounded mentions (2.43 vs. 1.57), and at the operating threshold $\delta = 1.5$, 82.0% of hallucinated mentions exceed the threshold compared with 51.1% of grounded mentions. Entropy is informative but imperfect: it identifies a large fraction of hallucination-prone mentions while triggering on grounded mentions. This result motivates using entropy to decide when to compute the local path, while retaining the relative-confidence comparison as a second safeguard before replacement.

Table 12: Mention-level entropy analysis on 100 CHAIR samples generated by LLaVA-v1.5-13B.

Statistic	Result
Object mentions (grounded / hallucinated)	709 / 122
Mention-level AUROC	0.735
Mean entropy (grounded / hallucinated)	1.57 / 2.43
Rate above $\delta = 1.5$ (grounded / hallucinated)	51.1% / 82.0%

Local-path usage. We distinguish the *trigger rate*, the fraction of decoding steps satisfying $H(p_t^{(1)}) > \delta$ for which the local forward pass is executed, from the *replacement rate*, the fraction of all steps at which the local prediction is finally adopted because it is more confident than the global prediction. Table 13 reports both quantities. The final column is the conditional acceptance rate, computed as replacement divided by trigger.

Table 13: Local-path usage across models and benchmarks. “Accept/trigger” is the fraction of triggered steps that adopt the local prediction.

Model	Benchmark	Trigger	Replacement	Accept/trigger
LLaVA-v1.5-13B	CHAIR-500	42.1%	24.9%	59.1%
LLaVA-v1.5-13B	AMBER-Gen	46.8%	27.4%	58.5%
MiniGPT-4	CHAIR-500	45.6%	25.0%	54.8%
MiniGPT-4	AMBER-Gen	48.8%	28.4%	58.2%

Across these settings, the local branch is computed for 42.1–48.8% of tokens but adopted for only 24.9–28.4% of all tokens. Equivalently, only 54.8–59.1% of triggered steps accept the local candidate. These statistics confirm selectivity at both stages: high entropy invokes an alternative prediction, but it does not automatically force stronger visual attention. They also help interpret the measured efficiency overhead in Appendix D, since the

global path is always evaluated whereas the local path is conditional.

E.3 Fixed-Parameter Transferability

To evaluate the robustness and transferability of GLVAC, we use a fixed parameter setting across different model sizes and datasets. Specifically, we set $(\gamma_1, \gamma_2) = (0.7, 1.4)$ for both LLaVA-v1.5-7B and LLaVA-v1.5-13B on CHAIR-500 and AMBER-Gen. Table 14 compares this setting with regular decoding.

Table 14: Fixed-parameter transferability on LLaVA-v1.5 using the shared setting $(\gamma_1, \gamma_2) = (0.7, 1.4)$ without model- or dataset-specific adjustment. The better result in each comparison is **bolded**.

CHAIR-500				
Model	Method	CHAIRs ↓	CHAIRi ↓	Recall ↑
LLaVA-7B	Regular	50.4	14.2	80.6
	GLVAC (shared)	47.2	13.7	81.3
LLaVA-13B	Regular	48.2	13.2	79.4
	GLVAC (shared)	47.0	12.8	80.8
AMBER-Gen				
Model	Method	Cover ↑	CHAIR ↓	Hal ↓ Cog ↓
LLaVA-7B	Regular	51.7	7.7	35.8 4.2
	GLVAC (shared)	51.3	7.3	33.4 3.9
LLaVA-13B	Regular	51.9	6.8	31.1 3.3
	GLVAC (shared)	52.9	6.6	32.1 3.3

The shared setting improves all three CHAIR metrics for both model sizes. On AMBER-Gen, it produces broadly comparable overall results to regular decoding across both model sizes. These results support transferability across model sizes and datasets.

E.4 Sensitivity Analysis of Attention Scaling Factors

Effect of Attention Scaling Factor γ_1 . We analyze the effect of the global-path scaling factor γ_1 by varying γ_1 while fixing $\gamma_2 = 1.1$ and the entropy threshold $\delta = 1.5$. The results on LLaVA-v1.5-13B using the CHAIR benchmark are shown in Figure 6. Note that $\gamma_1 = 1$ corresponds to the original, unmodified visual-attention strength. Overall, increasing γ_1 leads to higher hallucination rates: CHAIRs rises sharply as γ_1 increases toward 1.0, and CHAIRi shows an overall upward trend. Meanwhile, Recall improves rapidly from $\gamma_1 = 0.4$ to 0.6 and then largely saturates, peaking near $\gamma_1 = 0.7$ and changing slightly for $\gamma_1 \in [0.6, 0.9]$. These results suggest that a conservative global attention setting (e.g., 0.6-0.8) can substantially reduce hallucinations compared to the de-

Table 15: Ablation on excluding early layers when applying GLVAC (LLaVA-v1.5-7B on CHAIR). Excluding layers 0–1 improves recall and decreases CHAIRs.

Method	CHAIRs ↓	CHAIRi ↓	Recall ↑
Our method (excluding layers 0–1)	46.2	13.9	81.2
w/o early-layer exclusion	49.0	13.6	80.0

fault $\gamma_1 = 1$ with little impact on object coverage. However, when γ_1 becomes too small (e.g., 0.4), Recall drops markedly due to insufficient visual grounding. Overall, γ_1 controls the global attention regime and critically affects the trade-off between hallucination suppression and coverage.

Effect of Attention Scaling Factor γ_2 . We analyze the effect of the local-path scaling factor γ_2 by varying γ_2 while fixing $\gamma_1 = 0.7$ and $\delta = 1.5$. Results for LLaVA-v1.5-13B on the CHAIR benchmark are shown in Figure 7. The overall metrics exhibit only mild fluctuations over $\gamma_2 \in [0.8, 1.4]$, which is reasonable because the local correction path is activated only on a subset of high-entropy tokens. Importantly, however, the local path still provides consistent gains over the global-only baseline for $\gamma_2 \in [0.8, 1.3]$. For reference, using only a global scaling of $\gamma_1 = 0.7$ yields CHAIRi = 12.4, CHAIRs = 44.6, and Recall = 79.8. Across $\gamma_2 \in [0.8, 1.1]$, CHAIRi remains lower and Recall remains higher than this baseline, indicating that the local correction branch reliably improves both hallucination suppression and object coverage. This demonstrates that, although less sensitive than γ_1 , the local path plays a meaningful and stabilizing role, making GLVAC robust to the exact choice of γ_2 .

E.5 Selection of Layers

To determine which transformer layers should be modified by GLVAC, we first examine the layer-wise attention distribution of LLaVA-v1.5-13B. As shown in Figure 8, layers 0–1 allocate substantially stronger attention to visual tokens than deeper layers, and the visual attention begins to decay from layer 2 onward. This suggests that the first two layers function as a visual-ingestion stage that consolidates the raw visual representations before higher-level reasoning takes place.

Based on this observation, we apply GLVAC only from layer 2 and leave layers 0–1 untouched. To validate this design choice, we further conduct an ablation on LLaVA-v1.5-7B using the CHAIR benchmark. As shown in Table 15, removing the

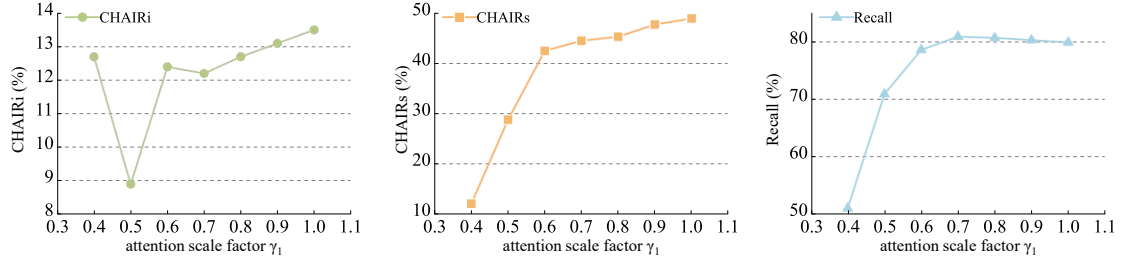


Figure 6: Sensitivity analysis of the global-path scaling factor γ_1 on LLaVA-v1.5-13B under the CHAIR benchmark ($\gamma_2 = 1.1$, $\delta = 1.5$; $\gamma_1 = 1$ is the original attention strength). Increasing γ_1 toward 1.0 increases CHAIRs and generally increases CHAIRi, while Recall saturates after $\gamma_1 = 0.6$ (peaking near $\gamma_1 = 0.7$). Thus, a conservative γ_1 (e.g., 0.6–0.8) suppresses hallucination with minimal Recall change.

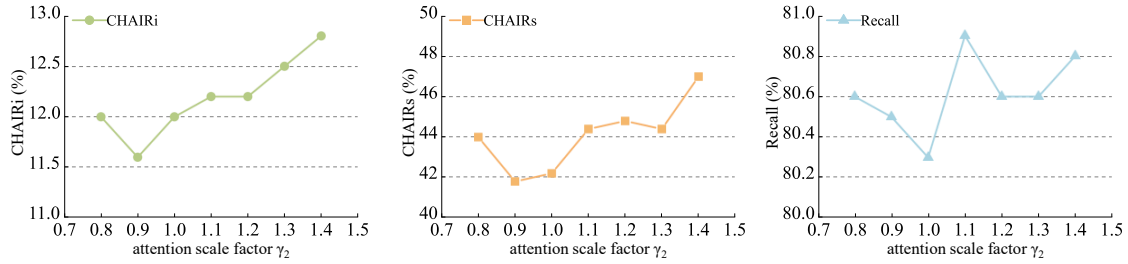


Figure 7: Sensitivity analysis of the local-path scaling factor γ_2 on LLaVA-v1.5-13B under the CHAIR benchmark ($\gamma_1 = 0.7$, $\delta = 1.5$). Although the metrics fluctuate only mildly across different γ_2 values, the local correction path consistently improves over the global-only baseline (CHAIRi = 12.4, CHAIRs = 44.6, Recall = 79.8) for $\gamma_2 \in [0.8, 1.1]$. This confirms that the local path provides meaningful token-level corrections.

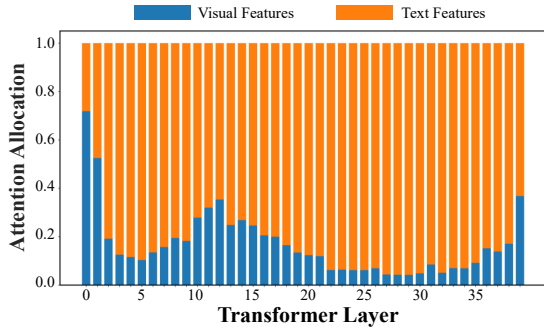


Figure 8: Layer-wise attention in LLaVA-v1.5-13B. Layers 0–1 allocate substantially stronger attention to visual tokens, while visual attention begins to decay from layer 2 onward. This motivates applying GLVAC only from layer 2 and leaving the first two layers unchanged.

early-layer exclusion consistently degrades both hallucination suppression and object recall. These results indicate that modifying the early layers disrupts the model’s initial visual-processing circuitry, and excluding them leads to more stable and effective visual grounding.

F AI Assistance Disclosure

AI assistants were used solely for language polishing and improving the clarity of the manuscript.

G Case Study

In this section, we present several case studies to illustrate the effectiveness of our approach. **Figure 9** and **Figure 10** show that GLVAC effectively mitigates object hallucination across diverse scenarios. **Figure 11** demonstrates that GLVAC further improves object recall. Finally, **Figure 12** and **Figure 13** illustrate that reducing visual attention can avoid errors made by the original model.

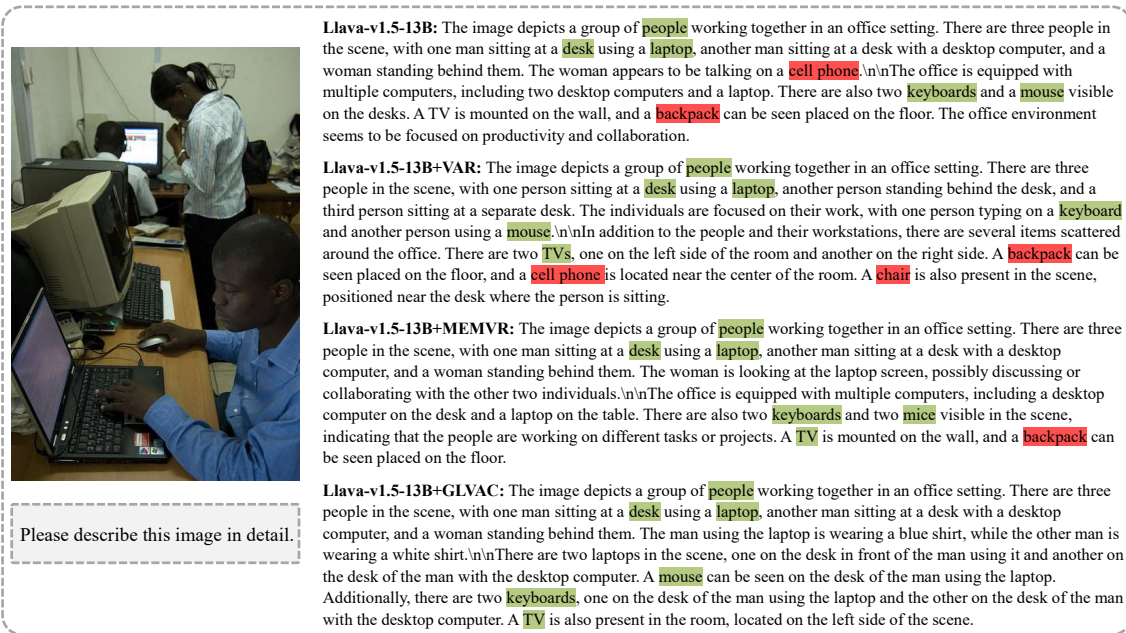


Figure 9: An illustration of GLVAC correcting hallucinations on CHAIR. ■ Green marks ground-truth objects. ■ Red marks hallucinated objects.

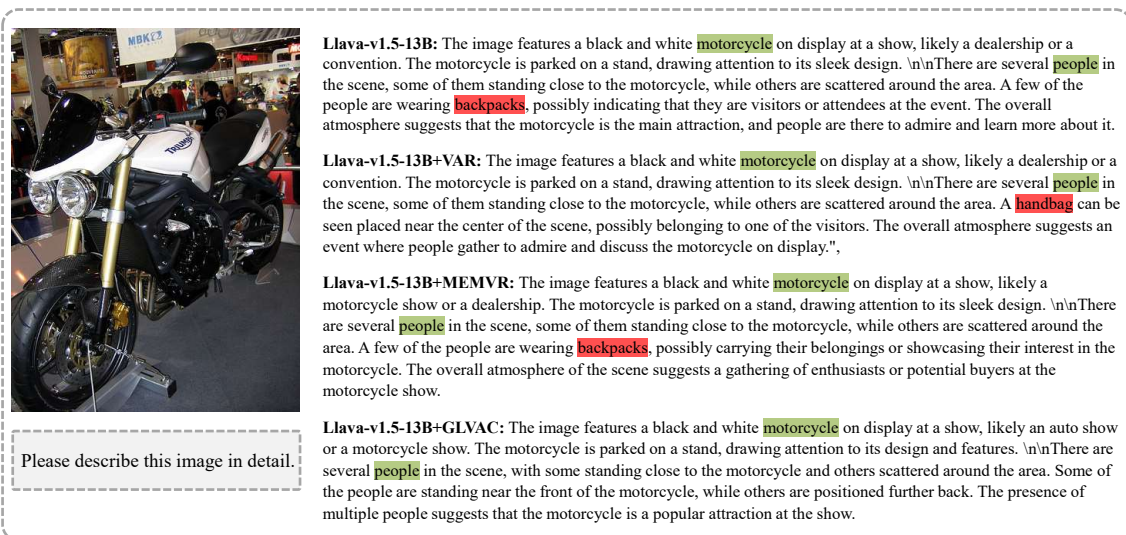


Figure 10: An illustration of GLVAC correcting hallucinations on CHAIR. ■ Green marks ground-truth objects. ■ Red marks hallucinated objects.



Figure 11: An illustration of GLVAC improving object recall on CHAIR. ■ Green marks ground-truth objects.

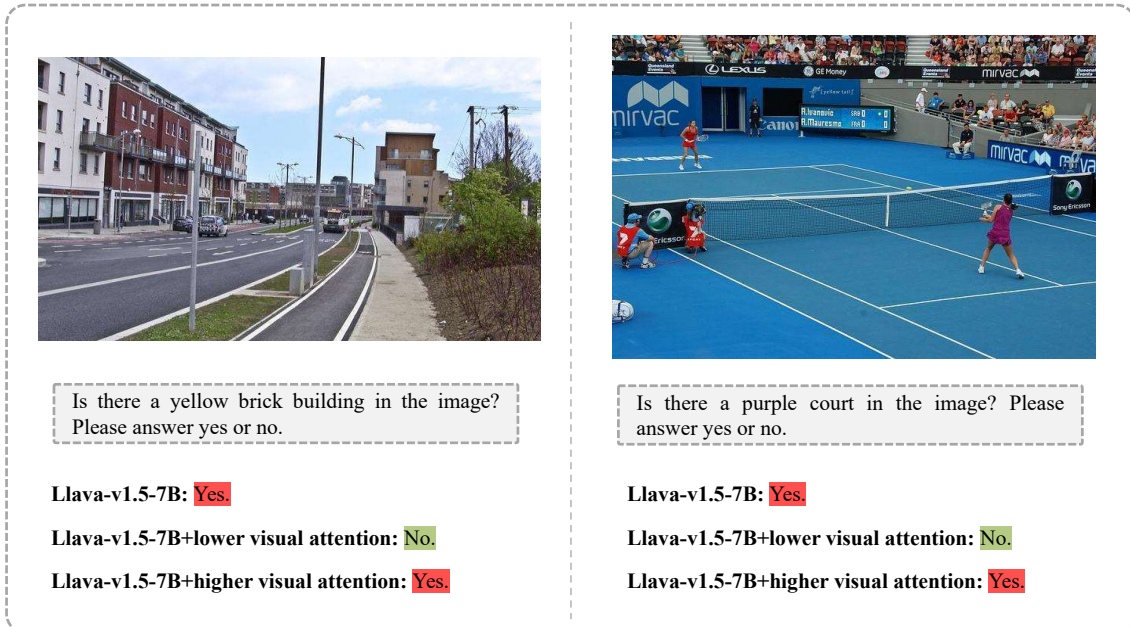


Figure 12: An illustration showing that reducing visual attention strength ($\gamma = 0.95$) can correct model predictions that are incorrect under the regular setting on MME. ■ Green marks correct response. ■ Red marks incorrect objects.



Figure 13: An illustration of reducing visual attention strength ($\gamma = 0.95$) can correct model predictions that are incorrect under the regular setting on MME. ■ Green marks correct response. ■ Red marks incorrect objects.