Looking Closer When Unsure: VIGOR, A Training-Free Framework for Mitigating Object Hallucination in Multimodal Large Language Models

Jiaxin Li[1], **Yixiang Chen**[3], **Lefan Lai**[4], **Geng Zhao**[5], **Xiaohan Long**[6], **Xuming Hu**[1,2†]

[1]The Hong Kong University of Science and Technology (Guangzhou)
[2]The Hong Kong University of Science and Technology
[3]The University of Sydney [4]The University of Auckland
[5]Medical University of Innsbruck [6]Shanghai Jiao Tong University
shelleylimeer1123@gmail.com xuminghu@hkust-gz.edu.cn

Abstract

Object hallucination remains a major challenge in Multimodal Large Language Models (MLLMs), even though they have achieved strong performance across a widely vision-language tasks. A key reason for object hallucination is limited fine-grained visual grounding provided by fixed-resolution visual encoding. In this paper, we propose **V**isual **I**nference **G**ating for **O**bject **R**efinement (VIGOR), a training-free inference-time framework for binary object verification. VIGOR adopts a selective coarse-to-fine pipeline that improves visual grounding without the overhead of always-on multi-view inspection. It first uses the global yes–no margin to estimate whether the available evidence is sufficient. When the evidence is insufficient, VIGOR performs local refinement on uncertain cases using a small set of overlapping crops. Finally, VIGOR combines global and local evidence with an asymmetric decision rule. Experiments on three benchmark datasets across five MLLM backbones show that VIGOR consistently outperforms strong training-free baselines, recovering missed fine-grained evidence while maintaining robust performance on challenging visual reasoning.



Figure 1: Comparison between a conventional MLLM and VIGOR for object prediction under uncertain visual evidence. Unlike the conventional model, which relies only on the global image, VIGOR triggers local multi-view inspection when the full-image prediction is uncertain, enabling recovery of missing evidence and a more grounded prediction.

1 Introduction

Multimodal Large Language Models (MLLMs) have demonstrated strong performance in cross-modal understanding and reasoning by jointly processing visual and textual information, achieving promising results across a wide range of vision-language tasks, including image captioning (Bucciarelli et al., 2024) and visual question answering (Kuang et al., 2025; Fang et al., 2025). However, MLLMs may sometimes generate descriptions of objects that are not actually present in the input image, a phenomenon known as object hallucination (Huang et al., 2025).

One important factor contributing to object hallucination in MLLMs is insufficient visual evidence (Li et al., 2023b). This reflects an upstream limitation jointly induced by statistical biases in the training data and the resolution bottleneck of the visual module (Leng et al., 2024), i.e., an architectural constraint that requires the input image to be downsampled and compressed into a limited set of visual tokens (Guo et al., 2024a). Consequently, the model is more likely to rely on learned priors than on perceptual grounding, leading to plausible but incorrect responses.

†Corresponding author.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 1463–1484
October 24-29, 2026 ©2026 Association for Computational Linguistics

Existing methods that address this visual-evidence limitation adopt diverse strategies. Pi et al. (2024) suppress pretraining bias through bootstrapped preference learning, which trains the model to prefer visually grounded responses over automatically generated but pretraining-biased negatives. However, such continual post-training methods mainly calibrate model behavior globally, rather than enabling the model to verify uncertain visual content in a specific test image. Decoding-time methods regulate token generation after visual representations have already been formed (Huang et al., 2024a; Leng et al., 2024). They can therefore suppress unsupported continuations, but cannot recover fine-grained visual evidence that was not reliably encoded in the first place. Cropping- or zooming-based methods (Shen et al., 2025; Park et al., 2025), by contrast, revisit local image regions to recover missing visual evidence. However, such methods often follow a fixed refinement procedure, applying additional visual inspection even when the initial global view may already be sufficient, resulting in unnecessary computational overhead.

Accordingly, our goal is to mitigate the practically important subset caused by insufficient visual evidence under fixed-resolution visual encoding. In light of this, to balance grounding accuracy and inference efficiency, we draw on the idea of active perception (Bajcsy et al., 2018), in which an initial global assessment is followed by selective local scrutiny when the initial assessment is uncertain. Inspired by this coarse-to-fine cognitive process, we propose **VIGOR** (Visual Inference Gating for Object Refinement), a training-free, plug-and-play framework. VIGOR first obtains an efficient initial prediction from a global visual pass, and uses the probability margin between positive and negative responses as an uncertainty signal. A gating mechanism then determines whether refinement is needed by comparing the estimated uncertainty with a selected threshold. If the estimated uncertainty remains below this threshold, VIGOR retains the global prediction to preserve efficiency. Otherwise, it performs multi-view refinement over a set of local crops to obtain finer-grained visual evidence for verifying the initial prediction. Finally, VIGOR integrates the local evidence with the initial global prediction and applies a consensus-based recalibration step to produce the final prediction. Figure 1 provides an overview of this coarse-to-fine inference process.

Experimental results show that VIGOR achieves strong performance in mitigating object hallucination across POPE-COCO, GQA-Binary, and MME, compared with seven baseline methods. Further analysis reveals that visual attention can shift across different VIGOR crops, indicating that different local views foreground distinct image regions and thereby expose complementary grounding cues.

Our contributions are summarized as follows:

- We propose VIGOR, a training-free, plug-and-play framework that selectively invokes local multi-view refinement based on perceptual uncertainty, enabling reliable fine-grained grounding with limited additional cost.
- Our results demonstrate the effectiveness of VIGOR across multiple MLLM backbones and challenging settings, and show that its adaptive cropping strategy provides complementary evidence by inducing attention shifts across views.

2 Related Work

It is widely acknowledged that object hallucination often arises when weak visual evidence causes the model to rely on language priors rather than perceptual grounding (Li et al., 2023b; Leng et al., 2024; Yang et al., 2025). Recent studies further suggest that limitations in visual grounding are a major source of object hallucination in MLLMs. For example, Guo et al. (2024b) argue that the fixed input resolution and limited number of visual tokens in MLLMs can weaken the representation of subtle local details. As a result, the lack of reliable visual evidence for fine-grained grounding is prone to induce hallucination. Li et al. (2025) discover that visual information is often progressively diluted during long-sequence generation, making one-shot global visual encoding gradually inadequate for reliable grounding. In addition, prior work also suggests that fine-grained distinctions not reliably captured by the vision encoder may lead to unstable grounding signals in MLLMs (Wang et al., 2025b).

To mitigate object hallucinations, some methods mitigate hallucination at inference time without additional training, primarily through decoding-level interventions such as distribution shaping or logit correction (e.g., OPERA (Huang et al., 2024a), VCD (Leng et al., 2024) and DeCo (Wang et al., 2025a)), rather than by conditionally acquiring local visual evidence. Such methods typically follow

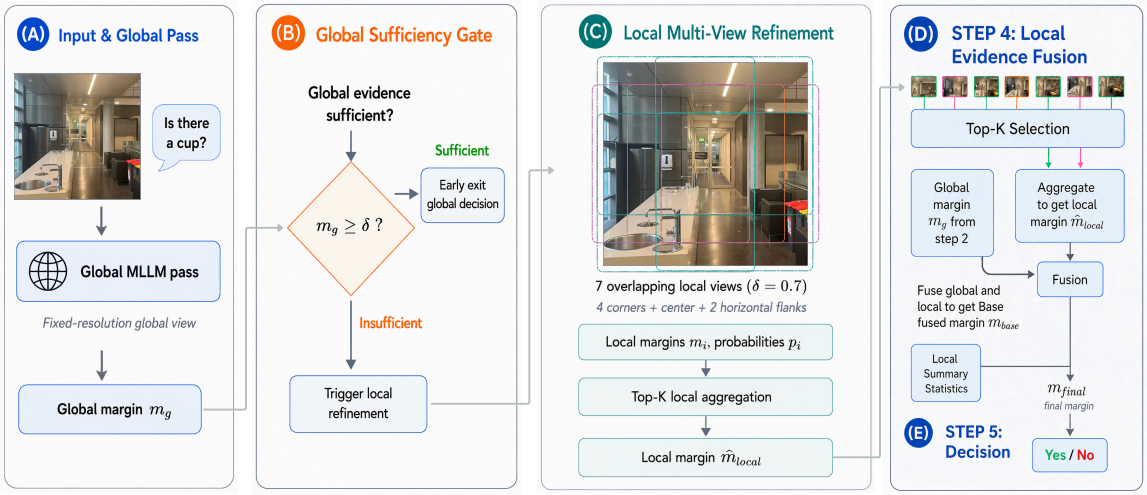


Figure 2: Overview of VIGOR. VIGOR first uses the global yes-no margin to decide whether the full-image evidence is sufficient. Routed cases are refined through seven overlapping local views, followed by top-K local aggregation, global-local evidence fusion, and asymmetric decision stabilization.

fixed refinement procedures, which may introduce unnecessary computation when global evidence is already sufficient. Existing resolution-scaling and multi-crop approaches typically rely on fixed, static slicing (e.g., Monkey (Li et al., 2024) and LLaVA-UHD (Guo et al., 2024b)), which improves fine-grained perception but often requires more visual tokens, increasing inference latency and potential context dilution by introducing excessive low-information visual slices. The limitations motivate a training-free approach that selectively acquires local visual evidence only when needed. VIGOR fills this gap through uncertainty-driven gating and conditional multi-view verification.

3 Method

We propose **VIGOR**, an inference-time framework for mitigating object hallucination in MLLMs for binary object verification. As shown in Figure 2, VIGOR uses selective coarse-to-fine inference: it first checks whether the global evidence is sufficient, applies local refinement only to routed cases through a small set of overlapping crops, and then combines global and local evidence with an asymmetric decision rule.

3.1 Task Setup

Let $\mathbf{I} \in [0,1]^{H\times W\times 3}$ denote an RGB image of height H and width W. Let $\mathbf{X}$ denote a textual query such as "Is there a dining table?" and let $\mathcal{M}_\theta$ denote an MLLM parameterized by θ.

We consider *binary object-existence verification*, where the response space is restricted to the two polar answers `yes` and `no`. Given an image view v and query $\mathbf{X}$, we append a binary-answer instruction and extract the model's final-step scores for the tokenizer-specific verbalizer tokens corresponding to `yes` and `no`. We denote these scores by

$$z_{yes}(v,\mathbf{X}) \in \mathbb{R}, \quad z_{no}(v,\mathbf{X}) \in \mathbb{R}. \tag{1}$$

We define the binary decision margin as $m(v,\mathbf{X}) = z_{yes}(v,\mathbf{X}) - z_{no}(v,\mathbf{X})$ where positive values favor `yes` and negative `no`. The corresponding pairwise positive response probability is

$$\begin{aligned} p(v,\mathbf{X}) &= \frac{\exp(z_{yes}(v,\mathbf{X}))}{\exp(z_{yes}(v,\mathbf{X})) + \exp(z_{no}(v,\mathbf{X}))} \\ &= \sigma(m(v,\mathbf{X}))\,, \end{aligned} \tag{2}$$

where $\sigma(u) = 1/(1+e^{-u})$ denotes the sigmoid function. For analysis, we further define an ambiguity score

$$\mathcal{U}(v,\mathbf{X}) = 1 - |2p(v,\mathbf{X}) - 1|\,, \tag{3}$$

which approaches 1 near the binary decision boundary and approaches 0 when the model strongly favors one response. We use $\mathcal{U}$ only as an analytical quantity; the routing rule implemented in VIGOR operates directly on the global yes–no margin.

3.2 Mechanistic Motivation: Resolution-Limited Evidence

VIGOR is motivated by a simple property of modern MLLMs: the input image is resized to a fixed encoder resolution before visual encoding. As a result, objects that occupy only a small spatial region in the input may receive weak visual representation after resizing (Guo et al., 2024b; Li et al., 2025;

Wang et al., 2025b).

Let the visual encoder consume an image resized to $R \times R$, where R denotes the encoder-side input resolution. For a queried object category o, let $\alpha_o(v) \in [0,1]$ denote the area ratio occupied by object o inside view v before resizing. Since the encoder resolution and patch size are fixed for a given MLLM, the number of visual tokens effectively allocated to the queried object is approximately proportional to its occupied area in the resized view, up to discretization, boundary, and saliency effects. We therefore use the following continuous proxy for *effective visual support*, $s_o(v) = \alpha_o(v)\, R^2$. This quantity is not intended as an exact token count; it serves as a simplified surrogate for how much of the encoder's spatial budget is effectively devoted to the queried object.

For the global view v_g, let $a_o \in [0,1]$ denote the area ratio of the queried object in the original image whenever the object is present. Then $\alpha_o(v_g) = a_o$ and $s_o(v_g) = a_o R^2$. When a_o is small, the queried object receives limited support in the resized global view. We refer to this as the *support-limited regime*.

Now consider a local crop view v_i spanning $\beta W \times \beta H$ pixels in the original image, where $\beta \in (0,1)$ is the crop side-length ratio. Its area ratio relative to the original image is $\rho = \beta^2$. If v_i still contains most of the queried object, then the object's area ratio inside the crop is magnified from a_o toward a_o/ρ, capped at 1. Hence,

$$\alpha_o(v_i) \approx \min\left\{1, \frac{a_o}{\rho}\right\},\ s_o(v_i) \approx \min\left\{1, \frac{a_o}{\rho}\right\} R^2. \tag{4}$$

In the support-limited regime where $a_o \ll \rho$, this simplifies to

$$\alpha_o(v_i) \approx \frac{a_o}{\rho}, \quad s_o(v_i) \approx \frac{a_o}{\rho} R^2 = \frac{s_o(v_g)}{\rho}. \tag{5}$$

Thus, local cropping increases effective visual support by approximately a factor of $1/\rho$ as long as the queried object remains inside the crop. In our implementation, $\beta = 0.7$, so $\rho = 0.49$ and the effective support can increase by approximately $2.04\times$ in this regime.

We make a mild monotonicity assumption: for ambiguous yet recoverable cases, increasing effective visual support tends to increase the magnitude of the binary decision margin. Formally, there exists a non-decreasing function $g : \mathbb{R}_{\geq 0} \to \mathbb{R}_{\geq 0}$ such that

$$\mathbb{E}[\,|m(v,\mathbf{X})| \mid s_o(v) = s\,] = g(s),\ g'(s) \geq 0. \tag{6}$$

The role of this assumption is motivational rather than identifiability-theoretic: VIGOR does not estimate object area or support explicitly at inference time. Instead, it uses the global yes–no margin as an operational proxy for evidential sufficiency. Under a fixed prompt and verbalizer protocol, a large positive global margin indicates stable affirmative support, while a small or negative margin indicates insufficient, conflicting, or weakly localized evidence. This makes the margin a cheap routing signal obtainable from the same global pass, without requiring explicit localization.

3.3 Stage I: Margin-Based Global Sufficiency Gating

Let v_g denote the resized full-image view used in the global pass, and let $m_g = m(v_g, \mathbf{X})$ denote the corresponding global margin. VIGOR uses a one-sided sufficiency gate to decide whether local refinement is needed:

$$\mathbb{G}(v_g, \mathbf{X}) = \mathbb{I}[m_g < \delta]\,, \tag{7}$$

where $\delta \in \mathbb{R}$ is a fixed trigger margin and $\mathbb{I}[\cdot]$ denotes the indicator function. When $\mathbb{G}(v_g, \mathbf{X}) = 0$, the global affirmative evidence is considered sufficient and VIGOR makes the prediction based on the global decision alone. When $\mathbb{G}(v_g, \mathbf{X}) = 1$, the example is routed to local refinement.

This gate is intentionally asymmetric. In binary object-existence verification, the dominant hallucination failure mode is unsupported affirmative prediction rather than unresolved negative uncertainty (Li et al., 2023b; Leng et al., 2024; Yang et al., 2025). Accordingly, VIGOR grants early exit to globally decisive affirmative cases, while routing all remaining cases—including weakly affirmative, ambiguous, and negative-leaning ones—to refinement. The key design principle is therefore *asymmetric sufficiency*: positive decisions must be supported strongly enough to justify stopping, whereas unresolved cases should remain eligible for evidence amplification.

3.4 Stage II: Local Support Amplification

3.4.1 Local view construction

When local-view refinement is triggered, VIGOR constructs a fixed set of overlapping local views

$$\mathcal{V}_{local} = \{v_1, \ldots, v_M\}, \quad M = 7. \tag{8}$$

Each local view uses the same side-length ratio $\beta = 0.7$. Specifically, the seven views consist of four corner crops, one center crop, and two horizon-

tal mid-flanking crops (Figure 2, Step 3). This layout increases the probability that a support-limited object receives enlarged representation in at least one crop, while avoiding the computational cost of denser tilings.

For each local view $v_i \in \mathcal{V}_{local}$, we compute its local margin and positive-response probability:

$$m_i = m(v_i, \mathbf{X}),\ p_i = p(v_i, \mathbf{X}),\ i = 1, \ldots, M. \tag{9}$$

Here, m_i measures the local yes–no decision margin of crop v_i, and p_i is the corresponding pairwise positive-response probability.

3.4.2 Existential local aggregation

Object existence is an *existential* task: a valid positive cue may appear in only one or a few informative crops. Averaging uniformly across all local views may therefore dilute sparse positive evidence with background-dominated crops. To reflect this asymmetry, VIGOR aggregates only the top-K local probabilities:

$$P_{local} = \frac{1}{K} \sum_{j=1}^{K} \mathrm{Top}_j(\{p_1, \ldots, p_M\}), \tag{10}$$

where $\mathrm{Top}_j(\cdot)$ denotes the j-th largest value and $K \leq M$ is the number of retained local views. We then map the aggregated local evidence back to a local margin:

$$\tilde{m}_{local} = \mathrm{logit}(P_{local}) = \log\left(\frac{P_{local}}{1 - P_{local}}\right). \tag{11}$$

Aggregating first in bounded probability space reduces sensitivity to extreme outlier crops, while top-K retention preserves the existential character of the task.

To match the implementation, we instantiate the local branch as a symmetric pseudo-score pair

$$z_{yes}^{(l)} = \frac{\tilde{m}_{local}}{2}, \quad z_{no}^{(l)} = -\frac{\tilde{m}_{local}}{2}, \tag{12}$$

so that $z_{yes}^{(l)} - z_{no}^{(l)} = \tilde{m}_{local}$. Let

$$z_{yes}^{(g)} = z_{yes}(v_g, \mathbf{X}), \quad z_{no}^{(g)} = z_{no}(v_g, \mathbf{X}) \tag{13}$$

denote the global scores. The global and local branches are then fused at the score level:

$$z_{yes}^{(f)} = w_g z_{yes}^{(g)} + w_l z_{yes}^{(l)},\ z_{no}^{(f)} = w_g z_{no}^{(g)} + w_l z_{no}^{(l)}, \tag{14}$$

which is equivalent, in margin space, to

$$m_{base} = z_{yes}^{(f)} - z_{no}^{(f)} = w_g m_g + w_l \tilde{m}_{local}. \tag{15}$$

We use weighted linear score fusion because both branches are produced by the same MLLM under the same binary verbalizer protocol, making them directly comparable as paired yes/no evidence. This yields a simple monotone combination rule whose effect remains transparent in margin space. The overall fusion and stabilization process is summarized in Steps 4–5 of Figure 2.

3.5 Stage III: Asymmetric Decision Stabilization

After fusion, VIGOR applies a lightweight asymmetric stabilization rule to routed cases. The purpose of this stage is not to replace the fused decision with a separate classifier, but to regularize hard cases in which local evidence is sparse, distributed, or weakly conflicting across crops.

We first compute the local summary statistics

$$\bar{m} = \frac{1}{M} \sum_{i=1}^{M} m_i, \quad m_{\max} = \max_{1 \leq i \leq M} m_i \tag{16}$$

and the affirmative and negative vote counts

$$V_+ = \sum_{i=1}^{M} \mathbb{I}[m_i > 0], \quad V_- = \sum_{i=1}^{M} \mathbb{I}[m_i \leq 0]. \tag{17}$$

Starting from

$$(z_{yes}^{(c)}, z_{no}^{(c)}) = (z_{yes}^{(f)}, z_{no}^{(f)}), \tag{18}$$

we instantiate four correction conditions:

$$\begin{aligned} \mathcal{C}_s &:= (m_{\max} > \eta_s) \wedge (V_+ \geq \nu_s), \\ \mathcal{C}_m &:= (V_+ \geq \nu_m) \vee (\bar{m} > \eta_m), \\ \mathcal{C}_l &:= (m_{\max} > \eta_l) \wedge (V_+ \geq \nu_l), \\ \mathcal{C}_n &:= (V_- \geq \nu_n) \wedge (m_{\max} < \eta_n), \end{aligned} \tag{19}$$

and apply exactly one of the following score corrections, evaluated in the listed priority order:

$$(z_{yes}^{(c)}, z_{no}^{(c)}) = \begin{cases} (z_{yes}^{(f)} + \Delta_s,\ z_{no}^{(f)}), & \text{if } \mathcal{C}_s, \\ (z_{yes}^{(f)} + \Delta_m,\ z_{no}^{(f)}), & \text{if } \mathcal{C}_m, \\ (z_{yes}^{(f)} + \Delta_l,\ z_{no}^{(f)}), & \text{if } \mathcal{C}_l, \\ (z_{yes}^{(f)},\ z_{no}^{(f)} + \Delta_n), & \text{if } \mathcal{C}_n, \\ (z_{yes}^{(f)},\ z_{no}^{(f)}), & \text{otherwise.} \end{cases} \tag{20}$$

These four conditions are not intended as a unique theoretically optimal partition of local evidence configurations. Rather, they operationalize four qualitatively distinct and task-relevant regimes: (i) a strong sparse positive cue, (ii) distributed moderate positive support, (iii) weaker but still recoverable positive tendency, and (iv) broad negative agreement in the absence of any strong positive crop. The first three branches provide progressively weaker positive recovery, while the negative branch is deliberately conservative.

Finally, to stabilize routed cases that remain close to the decision boundary after correction, we define the fallback condition

$$\mathcal{C}_f := \big(\mathbb{G}(v_g, \mathbf{X}) = 1\big) \wedge \big(|z_{yes}^{(c)} - z_{no}^{(c)}| < \epsilon\big) \wedge \big(V_+ \geq V_-\big), \tag{21}$$

and apply a low-margin fallback:

$$(z_{yes}^{(final)}, z_{no}^{(final)}) = \begin{cases} (z_{yes}^{(c)} + \Delta_f,\ z_{no}^{(c)}), & \text{if } \mathcal{C}_f, \\ (z_{yes}^{(c)},\ z_{no}^{(c)}), & \text{otherwise.} \end{cases} \tag{22}$$

The final margin and positive probability are then

$$m_{final} = z_{yes}^{(final)} - z_{no}^{(final)}, \quad p_{final} = \sigma(m_{final}), \tag{23}$$

and the model predicts `yes` if $m_{final} > 0$ and `no` otherwise. This asymmetric stabilization reflects the structure of binary object verification. A single highly informative crop, or a small set of mutually supportive crops, may be sufficient to recover a true positive. By contrast, negative correction should require broader agreement and the absence of any strong local positive evidence. In addition, the complexity analysis of VIGOR is provided in Appendix B.1.

4 Experiments

4.1 Experimental Setup

Dataset. We evaluate all methods on two authoritative benchmarks for object hallucination: POPE (Li et al., 2023b) and MME (Fu et al., 2024). For POPE, we focus on subsets sourced from MSCOCO (Lin et al., 2014) and GQA-Binary (Hudson and Manning, 2019). For MME, following the protocol from VCD (Leng et al., 2024), we select 5 subtasks from it. More details on the datasets and data splits are provided in Appendix C.1.

Baselines. Our baselines include four competitive training-free methods for mitigating hallucination: **VCD** (Leng et al., 2024), **DeCo** (Wang et al., 2025a), **OPERA** (Huang et al., 2024a), and **MASH-VLM** (Bae et al., 2025). In addition, we implement two post-hoc hallucination mitigation strategies that are widely used in industry: **SELF_CHECK** and **SLIDING**. These methods cover several key perspectives on object hallucination mitigation, e.g., signal augmentation and attention calibration. More details are provided in Appendix C.2. We select 5 mainstream MLLMs as the backbones for evaluation: LLaVA-1.5-7B/13B, Phi-3.5-vision-instruct, Qwen2.5-VL-7B, and Qwen-VL-Chat (reported in Appendix A.3). We plug both the baseline methods and VIGOR into the backbones and perform training-free inference without any parameter updates. Implementation details of VIGOR are in Appendix C.3.

4.2 Comprehensive Evaluations

Tables 1, 2 and Figure 3 show the results of comprehensive evaluations. We summarize main findings below. We provide additional hyperparameter sensitivity analyses in Appendix A.1 and stress tests under Gaussian blur and noise perturbations in Appendix A.5.

Efficient Object Verification, Operationalization, and Evidence Retrieval. On POPE-COCO (Table 1), VIGOR consistently outperforms the strongest baseline across diverse MLLMs by an average F1 margin of 1.21%, demonstrating the effectiveness of its selective coarse-to-fine design. On the challenging *Adversarial* split, strong language priors markedly hinder SLIDING and DeCo in recalling ambiguous objects: although their aggressive decoding suppresses false positives, it also expands the negative prediction space and dilutes sparse positive cues with background noise. By contrast, VIGOR grounds verification in localized multi-view evidence and thus avoids this failure mode. Likewise, although VCD algorithmically disentangles language priors, its weaker performance on harder settings suggests that intervention at textual decoding comes too late. VIGOR instead intervenes earlier, at the visual encoder, where overlapping crops recover high-frequency details to resolve spatial ambiguity before text generation begins.

Balancing Fine-grained Perception and Global Understanding. Overall, on MME (Table 2), VIGOR ranks first across all five evaluation metrics, outperforming the second-best method by an average margin of 1.47%. The results on individual subsets further show that, although VIGOR relies on intensive local crop extraction during refinement, it remains sufficiently robust to severe semantic drift on the Scene subset. Furthermore, consistent gains on *Position* and *Posters* confirm that dynamically bypassing globally obvious cases allows VIGOR to seamlessly balance fine-grained perception with holistic global understanding.

Preserving Core Visual Reasoning. Hallucination mitigation should not compromise intrinsic reasoning ability. On GQA-Binary (Table 2), which requires complex deduction and precise spatial awareness, baselines such as VCD that perform strongly on POPE-COCO suffer a marginal accuracy drop of 0.54%. The average 20.88% decline of OPERA and DeCo further suggests a limitation of symmetric logit penalties: they may disrupt the

Table 1: Comprehensive evaluation results on POPE-COCO. Metric: Accuracy and F1 (precision and recall are omitted for compact presentation). "Greedy" decoding is the "relative zero" with no algorithmic intervention. We conduct at least 5 runs and report the averaged results (score). The results that are significantly better under the approximate randomization test at $p \leq .05$ are underlined.

Base MLLM	Method	Random		Popular		Adversarial	
		Accuracy (%)	F1 Score (%)	Accuracy (%)	F1 Score (%)	Accuracy (%)	F1 Score (%)
LLaVA-1.5-7B (Liu et al., 2024a)	Greedy	85.87	83.87	84.90	82.95	82.93	81.09
	VCD	88.70	**88.73**	84.27	84.96	78.97	80.83
	DECO	85.40	86.59	81.77	83.79	75.47	79.36
	OPERA	87.37	85.98	85.77	84.48	83.47	82.41
	MASH_VLM	85.23	83.41	84.23	82.49	81.70	80.22
	SELF_CHECK	86.80	85.20	85.10	83.61	83.03	81.71
	SLIDING	80.17	75.38	79.80	75.04	78.97	74.19
	VIGOR	**89.40**	88.49	**87.57**	**86.76**	**85.10**	**84.51**
LLaVA-1.5-13B (Liu et al., 2024a)	Greedy	87.67	86.29	86.33	85.03	84.33	83.19
	VCD	90.37	**90.30**	85.80	86.36	82.27	83.51
	DECO	83.60	85.31	78.10	81.31	74.43	78.83
	OPERA	89.73	89.15	87.07	86.70	83.87	83.94
	MASH_VLM	87.17	85.96	85.83	84.73	83.50	82.66
	SELF_CHECK	87.43	85.95	86.47	85.03	84.67	83.35
	SLIDING	84.20	81.47	83.70	80.99	82.73	80.08
	VIGOR	**90.67**	90.09	**88.77**	**88.31**	**85.77**	**85.65**
Qwen2.5-VL-7B (Bai et al., 2025)	Greedy	88.33	86.94	87.40	86.04	86.63	85.33
	VCD	89.10	87.96	88.30	87.21	87.43	86.41
	DECO	86.60	87.94	79.30	82.52	72.53	78.05
	OPERA	89.53	89.03	86.30	86.11	84.23	84.37
	MASH_VLM	89.13	88.09	87.83	86.86	86.47	85.59
	SELF_CHECK	88.63	87.29	87.67	86.36	86.90	85.62
	SLIDING	82.90	79.41	82.30	78.84	82.07	78.62
	VIGOR	**90.50**	**89.67**	**89.13**	**88.36**	**87.67**	**86.98**
Phi-3.5-vision (Abdin et al., 2024)	Greedy	86.13	84.18	85.07	83.17	83.43	81.67
	VCD	88.93	**88.52**	86.20	86.06	83.60	83.84
	DECO	83.67	85.05	81.63	83.50	77.67	80.61
	OPERA	85.77	83.70	84.77	82.75	83.17	81.26
	MASH_VLM	85.67	83.88	84.90	83.17	83.43	81.83
	SELF_CHECK	85.07	82.68	84.23	81.88	82.70	80.47
	SLIDING	83.27	80.08	82.83	79.67	81.83	78.70
	VIGOR	**89.13**	88.26	**87.93**	**87.13**	**85.57**	**84.99**

Method	MME					GQA-Binary	
	Position	Posters	Count	Scene	Existence	Accuracy	F1
Greedy	86.67	92.52	85.00	86.00	**100.00**	85.56	86.31
DECO	61.67	65.99	53.33	78.00	83.33	69.32	78.38
VCD	85.00	92.52	**86.67**	86.50	**100.00**	85.02	85.82
MASH_VLM	81.67	91.16	81.67	85.00	**100.00**	85.09	85.98
OPERA	66.67	50.00	51.67	50.00	91.67	60.05	73.08
SELF_CHECK	85.00	91.84	**86.67**	84.75	**100.00**	84.90	85.71
SLIDING	81.67	90.14	81.67	86.25	93.33	82.83	82.92
VIGOR (Ours)	**90.00**	**92.86**	**86.67**	**88.50**	100.00	**85.79**	**87.30**

Table 2: Comprehensive evaluation results on MME and GQA-Binary Visual Reasoning. Metric: Accuracy (%) / F1-score (%); Base Model: Qwen2.5-VL. The results that are significantly better under the approximate randomization test at $p \leq .05$ are underlined.

model's latent multi-step reasoning. In contrast, VIGOR uses asymmetric decision stabilization, requiring broader evidence agreement before validating negative conclusions. This allows it to filter local noise while preserving the semantic coherence needed for complex visual reasoning.

Case Study. To further examine how VIGOR improves perceptual grounding, we present a case study on LLaVA-13B with the query, "Is there a backpack in the image?" As shown in Figure 3, the ground-truth answer is yes, yet the original MLLM over-relies on the global view and incorrectly predicts no. After crop-based refinement, VIGOR corrects the prediction to yes. Compared with the global attention, VIGOR's local crops attend more strongly and more precisely to the backpack region, suggesting that refinement recovers visual evidence missed in the global pass. The differing attention patterns across crops further indicate that distinct local views provide complementary cues.

4.3 Ablation Study

Figure 4 ablates the architectural components of VIGOR. The findings are: **(1)** Multi-view refinement drives the largest F1 gains (+2.69 over Global-Only). Extending from 4 to 7 crops further boosts Recall, proving that overcoming limited spatial resolution is the primary factor in mitigating hallucinations. **(2)** The gate is a pure efficiency switch that safely bypasses globally obvious queries to cut latency without sacrificing accuracy. This finding is consistent with our design. **(3)** Top-K plays

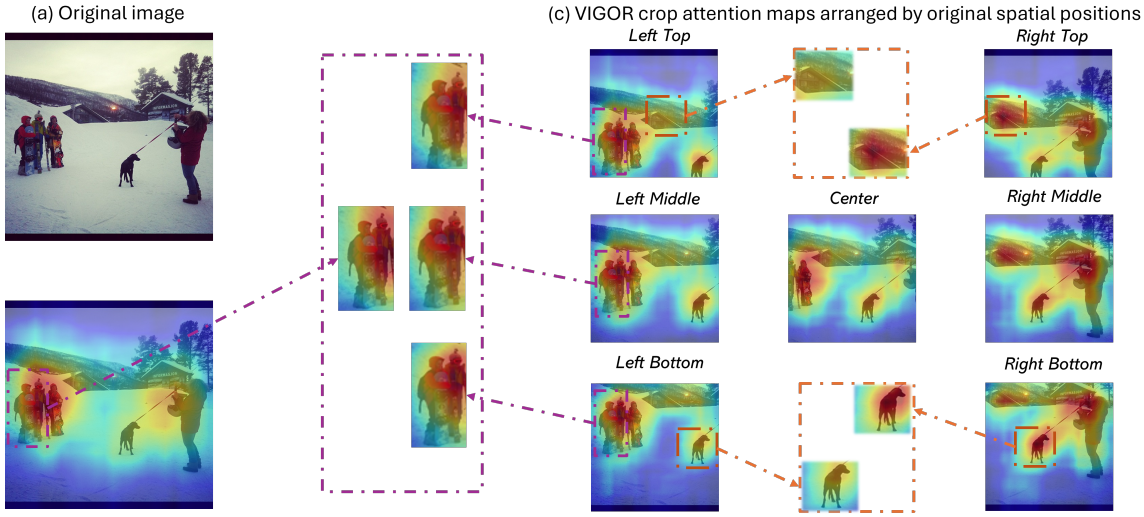


Figure 3: Case study: attention visualization on a hard instance. Base Model: LLaVA-1.5-13B; Question: Is there a backpack in the image; Ground truth: Yes; Response w/o VIGOR: No; Response w/ VIGOR: Yes. The purple dashed box highlights the comparison between the original global attention and VIGOR's crop-based attention over the backpack region. The orange dashed boxes show attention differences across crops.

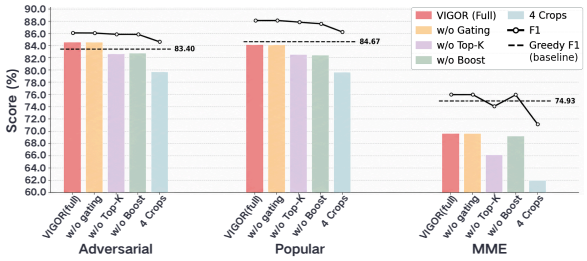


Figure 4: Ablation results on the Adversarial and Popular splits of POPE-COCO, and on MME, using LLaVA-1.5-13B as backbone model. Bars show recall, black line shows F1-score, and dashed line indicates Greedy F1 baseline.

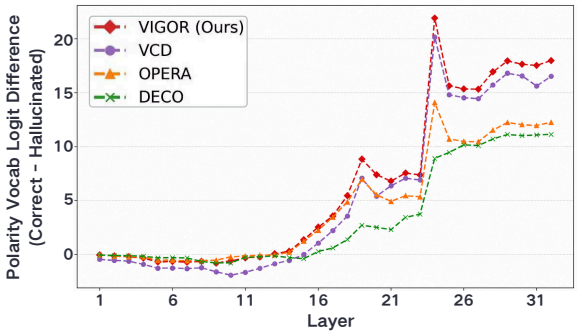


Figure 5: Layer-wise mechanistic localization analysis; Object: Residual stream; Base model: LLaVA-1.5-7B; Metric: Logit difference (correct-hallucinated); Dataset: POPE-COCO.

an incremental and positive role, as it can prevent background dilution for infrequent extreme long-tail ambiguities.

4.4 Mechanistic Investigation

To better understand object hallucination, we project the residual stream at each layer into the vocabulary space (Dar et al., 2023) to compute the difference between the summed-up logits of tokens from correct and hallucinated answer token subsets. In Figure 5, averaged across five datasets, VIGOR accumulates grounded visual evidence more effectively than the baselines throughout forward propagation. Notably, VIGOR starts to depress hallucination signals at earlier layers, part of the effect even emerges before reaching the layers for reasoning. We conjecture that the injected local views provide complementary fine-grained evidence that reshapes the intermediate representations early on, thereby improving the calibration of visually grounded concepts in the residual stream. More interpretability analyses are provided in Appendix B.2.

5 Conclusion

We propose VIGOR, a training-free inference framework that mitigates MLLMs object hallucination caused by limited visual resolution. VIGOR selectively invokes local crops through a sufficiency-based gating mechanism and aggregates evidence across views to recover fine-grained visual details that may be missed in the global view. Experiments show VIGOR effectively alleviates object hallucinations in MLLMs while maintaining strong performance on challenging visual reasoning tasks.

Limitations

We acknowledge that VIGOR has several limitations. First, VIGOR relies on a fixed crop prior for local visual refinement: it uses a predefined seven-view crop layout rather than dynamically selecting regions conditioned on each image-query pair. Although this design provides a simple and consistent basis for static query-context verification, it does not fully capture contextual drift in open-ended text generation. As a result, extending VIGOR to unconstrained auto-regressive decoding may introduce some redundant local inspections and slightly increase verification cost. Nevertheless, our benchmark results show that these limitations do not produce significant negative effects in the targeted settings, and they do not diminish VIGOR's advantages over baseline training-free hallucination mitigation methods.

Second, VIGOR currently operates in a binary verification framework. While this formulation is well aligned with hallucination mitigation under query-based evaluation, it does not fully capture the richer uncertainty structure of free-form generation.

Future work is discussed in Appendix E.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (Grant No. 62506318); the Guangdong Provincial Department of Education Project (Grant No. 2024KQNCX028); the Scientific Research Projects for Higher-Educational Institutions (Grant No. 2024312096), Education Bureau of Guangzhou Municipality; and the Guangzhou–HKUST(GZ) Joint Funding Program (Grant No. 2025A03J3957), Education Bureau of Guangzhou Municipality.

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, and 1 others. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Kyungho Bae, Jinhyung Kim, Sihaeng Lee, Soonyoung Lee, Gunhee Lee, and Jinwoo Choi. 2025. Mash-vlm: Mitigating action-scene hallucination in video-llms through disentangled spatial-temporal representations. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13744–13753.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Ruzena Bajcsy, Yiannis Aloimonos, and John K Tsotsos. 2018. Revisiting active perception. *Autonomous Robots*, 42(2):177–196.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, and 1 others. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

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.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024. Unified hallucination detection for multimodal large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3235–3252.

Guy Dar, Mor Geva, Ankit Gupta, and Jonathan Berant. 2023. Analyzing transformers in embedding space. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16124–16170.

Wenlong Fang, Qiaofeng Wu, Jing Chen, and Yun Xue. 2025. guided mllm reasoning: Enhancing mllm with knowledge and visual notes for visual question answering. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19597–19607.

Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. 2024. Multi-modal hallucination control by visual information grounding. In *Proceedings of the*

IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14303–14312.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yixun Qin, Mengdan Zhang, Lin Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and 1 others. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14427–14437.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, and 1 others. 2024. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 14375–14385.

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.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. 2024a. Llava-uhd: an lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision*, pages 390–406. Springer.

Zonghao Guo, Ruyi Xu, Yuan Yao, Junbo Cui, Zanlin Ni, Chunjiang Ge, Tat-Seng Chua, Zhiyuan Liu, and Gao Huang. 2024b. Llava-uhd: An lmm perceiving any aspect ratio and high-resolution images. In *European Conference on Computer Vision (ECCV)*, pages 390–406. Springer.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024a. 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*.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. 2024b. Visual hallucinations of multi-modal large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 9614–9631.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (CVPR)*, pages 6700–6709.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. 2025. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 25004–25014.

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.

Junyi Li, Xiaoxue Cheng, Wayne Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2023a. Halueval: A large-scale hallucination evaluation benchmark for large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 6449–6464.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 292–305.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024. Monkey: Image resolution and text label are important things for large multi-modal models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26763–26773.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N Metaxas. 2025. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. *arXiv preprint arXiv:2502.03628*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024b. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1906–1919.

Kyungmin Min, Minbeom Kim, Kang-il Lee, Dongryeol Lee, and Kyomin Jung. 2025. Mitigating hallucinations in large vision-language models via summary-guided decoding. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4183–4198.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Joonhyung Park, Peng Tang, Sagnik Das, Srikar Appalaraju, Kunwar Yashraj Singh, R Manmatha, and Shabnam Ghadar. 2025. R-vlm: Region-aware vision language model for precise gui grounding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 9669–9685.

Renjie Pi, Tianyang Han, Wei Xiong, Jipeng Zhang, Runtao Liu, Rui Pan, and Tong Zhang. 2024. Strengthening multimodal large language model with bootstrapped preference optimization. In *European Conference on Computer Vision*, pages 382–398. Springer.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045.

Haozhan Shen, Kangjia Zhao, Tiancheng Zhao, Ruochen Xu, Zilun Zhang, Mingwei Zhu, and Jianwei Yin. 2025. Zoomeye: Enhancing multimodal llms with human-like zooming capabilities through tree-based image exploration. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 6613–6629.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, and 1 others. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 13088–13110.

Chenxi Wang, Xiang Chen, Ningyu Zhang, Bozhong Tian, Haoming Xu, Shumin Deng, and Huajun Chen. 2025a. MLLM can see? dynamic correction decoding for hallucination mitigation. In *International Conference on Learning Representations (ICLR)*.

Weihang Wang, Xinhao Li, Ziyue Wang, Yan Pang, Jielei Zhang, Peiyi Li, Qiang Zhang, and Longwen Gao. 2025b. Diving into mitigating hallucinations from a vision perspective for large vision-language models. *arXiv preprint arXiv:2509.13836*.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15840–15853.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. 2025. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 25543–25551.

Tianyun Yang, Ziniu Li, Juan Cao, and Chang Xu. 2025. Understanding and mitigating hallucination in large vision-language models via modular attribution and intervention. In *The Thirteenth International Conference on Learning Representations*.

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. 2024. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11991–12011.

A Additional Experiments

A.1 Sensitivity Analysis of Hyperparameters

A.1.1 Sensitivity to the Gating Threshold

We examine the effect of the gating threshold on local-refinement frequency, computational efficiency, and model performance. As shown in

Gate Mode	Accuracy (%)	F1 Score (%)	Trigger Rate (%)	Relative FLOPs (× Global)	Wall Time (s)	Throughput (samples/s)	Peak Memory (GB)
Global only	86.70	85.41	0.00	1.00	460.15	6.52	15.71
Always local	87.63	86.99	100.00	8.00	2410.14	1.24	15.71
$\delta = 2.5$	87.63	86.99	68.53	5.80	1791.31	1.67	15.71
$\delta = 3.0$	87.63	86.99	70.00	5.90	1801.61	1.67	15.71
$\delta = 3.6$ **(Ours)**	**87.63**	**86.99**	**72.33**	**6.06**	**1847.04**	**1.62**	**15.71**
$\delta = 4.0$	87.63	86.99	73.53	6.15	1869.29	1.60	15.71
$\delta = 5.0$	87.63	86.99	77.00	6.39	1937.39	1.55	15.71

Table 3: Sensitivity analysis of the gating threshold δ on Qwen2.5-VL-Chat across the POPE-Adversarial split. The threshold mainly controls the trigger rate and computational cost, while accuracy and F1 score remain stable across the tested threshold values.

Table 3, a lower gating threshold results in fewer local-refinement triggers and therefore improves image-processing efficiency. Conversely, a higher gating threshold increases the frequency of local refinement, leading to higher computational cost and lower processing efficiency.

Figure 6 shows that the model maintains relatively stable performance across the tested gating thresholds, suggesting that the gating threshold primarily regulates the balance between computational efficiency and local-refinement frequency, rather than substantially affecting prediction quality. To balance performance and efficiency while avoiding an overly dataset-specific threshold, we choose the elbow point of the curve as the final setting. Therefore, we set the gating threshold to 3.6 in VIGOR.

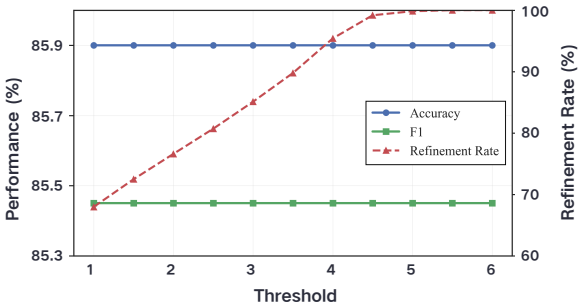


Figure 6: Effect of threshold δ on performance and refinement rate for LLaVA-1.5-13B. Increasing δ raises the proportion of samples sent to local refinement, while accuracy and F1 remain largely stable.

A.1.2 Sensitivity Analysis of Top-K Aggregation

The hyperparameter K denotes the number of selected local crops used in the final aggregation stage. Specifically, after VIGOR evaluates multiple local crops, it selects the top-K crops with the strongest local evidence and aggregates them with the global prediction. Therefore, K determines how much local evidence contributes to the final decision. If K is too small, such as $K = 1$, the method may be dominated by a single spurious crop. If K is too large, additional weak or irrelevant local evidence may dilute useful positive evidence. Consequently, selecting $K = 2$ allows VIGOR to preserve sparse positive evidence while avoiding both excessive dilution and full dependence on a single spurious crop. We further conduct a sensitivity analysis on K. As shown in Table 4, $K = 2$ yields the highest accuracy and F1 score among the tested strategies, while maintaining a competitive precision-recall balance.

Strategy	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Mean	87.27	92.93	80.67	86.37
Max	87.57	91.83	82.37	86.83
Top-2 (Ours)	**87.63**	**91.78**	**82.67**	**86.99**
Top-3	87.37	91.99	81.87	86.63
Top-4	87.47	92.45	81.60	86.69
Top-5	87.43	92.70	81.27	86.61

Table 4: Sensitivity analysis of the top-K aggregation strategy on Qwen2.5-VL-Chat across the POPE-Adversarial split. Top-2 aggregation achieves the best overall performance among the tested strategies, with the highest accuracy and F1 score while maintaining a competitive precision-recall trade-off.

A.1.3 Experiments of Crop Configurations

We first conduct a sensitivity analysis on crop ratio β. For crop ratio, given the spatial coverage and shape of our crop design, β should fall within an optimal range. If $\beta \leq 0.5$, small objects located near the vertical centerline of the image may be truncated or overlooked by the local views. In contrast, if $\beta = 1$, each crop becomes equivalent to the full image, causing the method to degenerate into repeatedly querying the global image. Therefore, β should be selected from the interval $(0.5, 1)$ and should not be too close to either boundary. Accordingly, we evaluate crop sizes within this interval.

The results in Table 5 show that $\beta = 0.7$ achieves the best performance among the tested settings.

Crop Ratio (β)	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
0.60	87.60	91.78	82.60	86.95
0.70 (Ours)	**87.63**	**91.78**	**82.67**	**86.99**
0.80	87.50	92.52	81.60	86.72
0.90	87.43	92.51	81.47	86.64

Table 5: Sensitivity analysis of the crop ratio in VIGOR on Qwen2.5-VL-Chat across the POPE-Adversarial split. Accuracy, precision, recall, and F1 score remain stable across the tested crop ratios, with the default ratio of 0.70 achieving the best overall performance.

We then conduct a sensitivity analysis on the number of crops used in VIGOR. Table 6 reports corresponding results on LLaVA-1.5-13B. As the number of local views increases from 1 to 7, both accuracy and F1 score improve steadily, suggesting that incorporating more local crops helps the model recover finer-grained visual evidence that may be missed in the global view. At the same time, relative latency increases progressively, reflecting the additional computational cost of processing more crops. When the number of local views is further increased to 9, performance exhibits a notable decline despite the higher latency, indicating that excessive local views may introduce redundancy or noise rather than useful complementary information. These results show that a moderate number of local views is most effective for LLaVA-1.5-13B, with $N = 7$ providing the best trade-off between performance and efficiency.

Finally, to examine the sensitivity of VIGOR to crop position, we examine whether different spatial arrangements of local views affect the overall performance. We evaluate different crop-position configurations, including a grid-based strategy. The results in Table 7 show that our proposed crop layout achieves the highest accuracy among the tested configurations.

A.1.4 Sensitivity Analysis of Fusion Weights

The fusion weights control how global and local evidence is integrated into the final decision. Specifically, w_g denotes the weight assigned to the global prediction, while w_l denotes the weight assigned to the aggregated local evidence. These weights help maintain decision stability by encouraging VIGOR to use both global context and local visual details, rather than relying entirely on a single source of evidence. To evaluate the impact of different weight balances, we conduct a sensitivity analysis across various configurations. As shown in Table 8, the default setting achieves the best performance among the tested configurations.

A.1.5 Sensitivity Analysis of the Asymmetric Stabilization Parameters

The asymmetric stabilization module contains several scalar parameters that determine how strongly local evidence can modify the global prediction. These parameters regulate either the minimum local margin required for a crop to be considered reliable or the minimum number of local crops required to support a correction. Evaluating each scalar parameter independently would introduce a large search space and make the resulting analysis difficult to interpret. Therefore, we group these parameters into three interpretable preset configurations: Conservative, Default, and Aggressive.

As expected, the Default configuration is adopted as the main setting in VIGOR. The Conservative configuration applies stricter correction criteria by requiring stronger local agreement before allowing positive recovery. In contrast, the Aggressive configuration increases the strength of positive recovery and lowers the evidence-margin thresholds, while still requiring support from multiple local views. This design allows us to examine whether VIGOR is sensitive to weaker or stronger stabilization behavior. Table 9 summarizes the three asymmetric stabilization configurations.

Table 10 reports their performance on the POPE-Adversarial split using Qwen2.5-VL-Chat. The Conservative configuration achieves high precision but lower recall, indicating that overly cautious correction may suppress useful positive evidence. The Aggressive configuration also maintains high precision but does not improve recall or F1 score, suggesting that stronger correction does not necessarily improve overall performance. In comparison, the Default configuration yields the highest accuracy and F1 score, providing the most favorable empirical balance between reliable correction and positive-evidence recovery.

A.2 Ablation Study of VIGOR on LLaVA-1.5-7B

Figure 7 presents an ablation study of VIGOR on LLaVA-1.5-7B over POPE-Adversarial, POPE-Popular, and MME. Overall, VIGOR achieves the best or near-best performance in all three settings. Removing the gating mechanism produces almost

Number of Crops	Crop Positions	Accuracy (%)	F1 Score (%)	Rel. Latency
1	Global image only	84.27	83.11	1.00×
3	Middle-left; Center; Middle-right	82.43	80.94	3.70×
5	Top-left; Top-right; Center; Bottom-left; Bottom-right;	85.53	84.78	5.60×
7 (Ours)	**Top-left; Top-right;Middle-left; Center; Middle-right; Bottom-left; Bottom-right;**	**85.90**	**85.45**	**7.40×**

Table 6: Sensitivity analysis of the number of crops in VIGOR on LLaVA-1.5-13B. The single-crop setting uses only the global image without local views. Increasing the number of crops improves performance from three to seven crops, and the seven-crop configuration achieves the best accuracy and F1 score among the tested settings.

Method	Crop Positions	Accuracy (%)	F1 Score (%)
VIGOR (Ours)	**Top-left; Top-right; Bottom-left; Bottom-right; Center; Middle-left; Middle-right**	**87.90**	87.26
Vertical	Top-left; Top-center; Top-right; Center; Bottom-left; Bottom-center; Bottom-right;	87.80	87.17
Main Diagonal	Top-left; Top-center; Middle-left; Center; Middle-right; Bottom-right; Bottom-center;	87.60	86.93
Anti Diagonal	Top-center; Top-right; Middle-left; Center; Middle-right; Bottom-left; Bottom-center;	87.57	86.83
Full Grid	Top-left; Top-center; Top-right; Middle-left; Center; Middle-right; Bottom-left; Bottom-center; Bottom-right	87.87	**87.33**

Table 7: Sensitivity analysis of crop positions in VIGOR on Qwen2.5-VL-Chat across the POPE-Adversarial split. VIGOR uses the seven-position layout with four corners, the center, and the left/right middle regions, achieving the highest Accuracy and a competitive F1 score.

Setting (w_g, w_l)	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
(0.20, 0.80)	87.63	91.35	83.13	87.05
(0.35, 0.65)	87.56	91.58	82.73	86.93
(0.52, 0.48) (Ours)	**87.70**	**91.92**	**82.66**	**87.05**
(0.65, 0.35)	87.53	91.89	82.33	86.85
(0.80, 0.20)	87.50	92.26	81.87	86.75

Table 8: Sensitivity analysis of fusion weights on Qwen2.5-VL-Chat across the POPE-Adversarial split. The selected setting, $w_g = 0.52$ and $w_l = 0.48$, achieves the highest accuracy and F1 score among the tested configurations while maintaining a competitive F1 score.

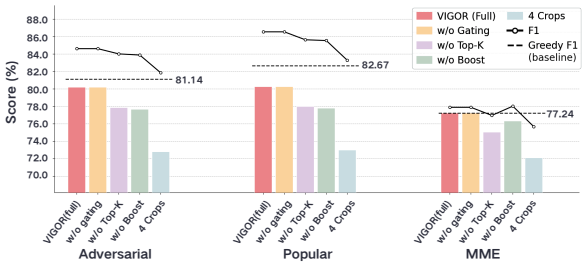


Figure 7: Ablation results of VIGOR on LLaVA-1.5-7B across POPE-Adversarial, POPE-Popular, and MME. Bars denote recall, the solid line denotes F1-score, and the dashed line indicates the Greedy baseline F1.

no change in recall or F1, suggesting that the gating mechanism improves efficiency with minimal performance loss. By contrast, removing Top-K fusion or boosting consistently reduces performance, particularly on MME, highlighting the importance of selectively aggregating informative local views. The 4-crop variant shows the largest decline across datasets, especially on MME, indicating that insufficient local coverage substantially weakens the recovery of missed positive evidence. These results suggest that VIGOR's effectiveness primarily arises from sufficiently dense multi-view inspection together with selective evidence aggregation.

A.3 Qwen-VL-Chat Results on POPE

We also evaluate VIGOR on POPE using Qwen-VL-Chat. Table 11 shows that VIGOR remains consistently effective on this backbone, achieving the best performance on the more challenging splits, including Popular in accuracy (88.07) and Adversarial in both accuracy (85.60) and F1 (84.95). Although VIGOR does not outperform all baselines on the Random split, it remains competitive overall. These findings further suggest that VIGOR is particularly beneficial in harder settings, where stronger perceptual grounding is needed to counter misleading or insufficient visual evidence.

A.4 Variance and Stability Analysis

In this part, we examine whether all methods can maintain stable hallucination suppression under challenging conditions, where visual evidence is ambiguous or more fine-grained or edged objects are present. Specifically, on POPE-COCO-Adversarial, we vary the hyperparameters "temperature," "top-p," and "top-k," and also change the random seeds, resulting in more than 50 repeated trials for each group. We report the averages across settings and seeds in Figure 8. The results show

Preset	Strong Positive (η, ν, Δ)	Medium Positive (η, ν, Δ)	Light Positive (η, ν, Δ)	Negative (η, ν, Δ)	Fallback (ϵ, Δ_f)
Conservative	$(1.00, 4, 1.60)$	$(0.30, 5, 1.00)$	$(0.40, 3, 0.60)$	$(-0.80, 5, 1.20)$	$(0.20, 0.15)$
Default (Ours)	$(0.80, 3, 2.00)$	$(0.20, 4, 1.40)$	$(0.25, 2, 0.90)$	$(-1.00, 5, 1.00)$	$(0.30, 0.25)$
Aggressive	$(1.10, 5, 1.20)$	$(0.35, 6, 0.80)$	$(0.45, 4, 0.40)$	$(-0.70, 4, 1.40)$	$(0.15, 0.05)$

Table 9: Configurations of asymmetric stabilization parameters. The Default setting is used as the main configuration of VIGOR, while the Conservative and Aggressive settings provide policy-level variants for evaluating the sensitivity of the stabilization rules.

Preset	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)
Conservative	87.63	92.93	81.47	86.82
Default (Ours)	**87.70**	**91.92**	**82.67**	**87.05**
Aggressive	87.53	93.04	81.13	86.68

Table 10: Experimental results of asymmetric stabilization parameters on Qwen2.5-VL-Chat across the POPE-Adversarial split. The Default setting achieves the highest accuracy, recall, and F1 score among the tested configurations, while the Conservative and Aggressive settings yield slightly higher precision but lower recall and F1 score.

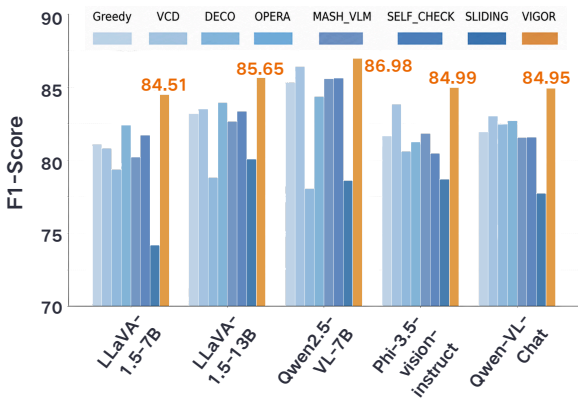


Figure 8: Performance variance analysis on POPE-COCO-Adversarial split.

that, compared with strong baselines such as VCD, VIGOR still exhibits stronger stability, with an advantage of about 15%, despite its local multi-view design being inherently more sensitive to randomness. Moreover, we find that VIGOR's stability advantage generally becomes larger as the backbone model is more prone to object hallucination and as its parameter scale becomes smaller.

A.5 Stress Testing

To comprehensively assess VIGOR's robustness to corrupted visual inputs, we conduct stress testing on the challenging POPE-COCO *Adversarial* split using LLaVA-1.5-13B. We add visual perturbations, including Gaussian Blur ($\sigma = 2.0$) and Gaussian Noise ($\sigma = 25$), into the images, to create a "language prior trap" for the methods.

As shown in Table 12, although image corruptions inevitably reduce the absolute performance of all methods, VIGOR consistently achieves a higher advantage regarding Accuracy and F1 scores under both perturbation types. In contrast, decoding-stage interventions such as DeCo and VCD suffer substantial drops in Precision, to around 66% and 77%, respectively. For DeCo, this degradation leads to deceptively high Recall at the cost of unacceptably high false positive rates. These results reveal a key weakness: when global visual cues are corrupted, vanilla models and decoding-only baselines tend to hallucinate by over-relying on language priors. For this reason, we argue that, VIGOR avoids this failure through local multi-view refinement. Even when global semantic features are severely degraded by noise or blur, VIGOR can zoom into overlapping local crops and verify object existence from the remaining high-frequency details. This explicit spatial anchoring stabilizes the verification process and makes the model more resistant to adversarial prompts that induce prior-driven predictions.

A.6 Additional Metrics

Table 13 further examines model behavior from the perspective of precision and recall. Across all backbones, VIGOR exhibits a consistently balanced profile over the three POPE splits. Relative to the original decoding results, it maintains strong precision while achieving clear improvements in recall, indicating that its gains are not merely the result of making the model more conservative. These results suggest that VIGOR enhances visual grounding in a manner that simultaneously supports reliable suppression of hallucinated objects and more accurate identification of objects that are actually present.

Table 11: Performance of Qwen-VL-Chat with different training-free hallucination mitigation methods on the POPE benchmark. Results are reported under the Random, Popular, and Adversarial settings using Accuracy (%) and F1-score (%).

Base MLLM	Method	Random		Popular		Adversarial	
		Accuracy (%)	F1 Score (%)	Accuracy (%)	F1 Score (%)	Accuracy (%)	F1 Score (%)
Qwen-VL-Chat (Bai et al., 2023)	NONE	86.47	84.55	85.23	83.38	83.67	81.95
	VCD	**90.80**	**90.54**	85.87	86.18	82.00	83.04
	DECO	90.00	90.10	87.43	**87.87**	80.67	82.47
	OPERA	87.97	86.65	86.30	85.08	83.67	82.70
	MASH_VLM	86.20	84.46	84.90	83.24	83.00	81.55
	SELF_CHECK	85.97	83.88	84.87	82.83	83.47	81.58
	SLIDING	82.33	78.68	81.87	78.24	81.30	77.73
	VIGOR	89.47	88.48	**88.07**	87.15	**85.60**	**84.95**

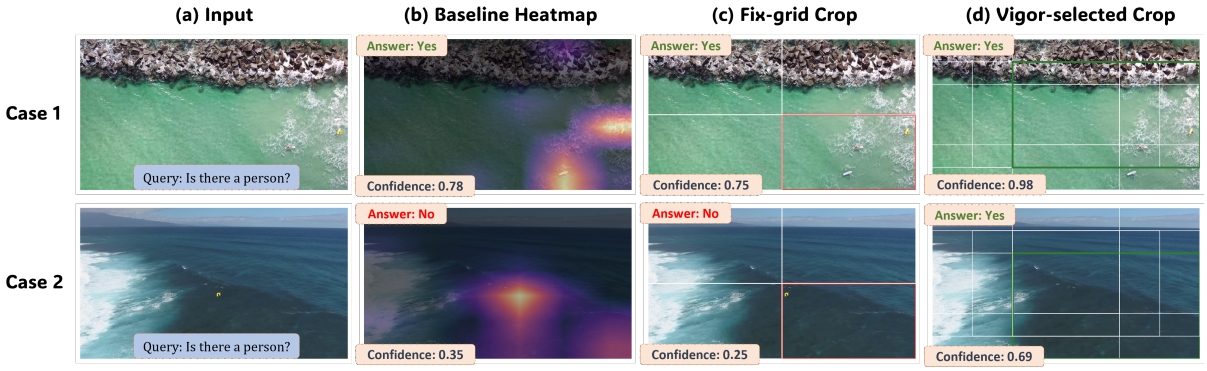


Figure 9: Qualitative comparison between the original model and VIGOR on the TinyPerson benchmark using Qwen2.5-VL-Chat. Each row shows one example. From left to right, the columns present the original image with the query, the baseline attention heatmap, the baseline grid-based crop layout, and the VIGOR crop layout. In the top example, VIGOR better preserves the tiny target within the selected local region and increases the model confidence. In the bottom example, the baseline model fails to detect the person, whereas VIGOR provides a more suitable local view and corrects the false negative prediction.

Table 12: Stress testing under visual perturbations (Gaussian Blur and Noise) on the POPE-COCO Adversarial split. All experiments are conducted using LLaVA-1.5-13B as the backbone.

Method	**Accuracy (%)**	**Precision (%)**	**Recall (%)**	**F1 Score (%)**
Setting I: Gaussian Blur				
Greedy	83.00	90.11	74.13	81.35
SELF_CHECK	82.47	**90.86**	72.20	80.46
DECO	73.63	66.68	**94.47**	78.18
VCD	81.27	77.65	87.80	82.42
VIGOR (Ours)	**85.10**	87.05	82.47	**84.70**
Setting II: Gaussian Noise				
Greedy	83.27	89.73	75.13	81.79
SELF_CHECK	83.40	**90.73**	74.40	81.76
DECO	73.40	66.39	**94.80**	78.09
VCD	81.63	77.90	88.33	82.79
VIGOR (Ours)	**84.70**	86.02	82.87	**84.41**

A.7 Qualitative Analysis

To further intuitively illustrate the effectiveness of the VIGOR layout under extreme conditions, Figure 9 presents a qualitative visualization on the TinyPerson benchmark using Qwen2.5-VL-Chat. The figure contains two representative examples, where each row corresponds to one query asking whether a person is present. These examples highlight the key challenges of TinyPerson, including vast maritime backgrounds, extremely small human targets, and targets that often appear near image boundaries.

In the first example, shown in the top row of Figure 9, the target person is located on the far right side of the shoreline. The global view predicts "YES" with a moderate confidence score of 0.78. However, the baseline grid-based crop layout places the tiny target close to a rigid crop boundary, which limits the visual evidence available within a single local region and slightly reduces the confidence to 0.75. In contrast, VIGOR selects a local region that better encloses the tiny person and its surrounding context, increasing the confidence to 0.98.

The second example, shown in the bottom row of Figure 9, presents a more challenging case in which the person is located in the open sea and oc-

Table 13: Precision and recall results on POPE across three evaluation splits. Higher precision indicates fewer false positives, while higher recall reflects stronger ability to recover ground-truth positive instances.

Base MLLM	Method	Random		Popular		Adversarial	
		Precision (%)	Recall (%)	Precision (%)	Recall (%)	Precision (%)	Recall (%)
LLaVA-1.5-7B	Greedy	97.70	73.47	95.25	73.47	90.89	73.20
	VCD	88.52	88.93	81.38	88.87	74.26	88.67
	DECO	80.07	**94.27**	75.41	**94.27**	68.49	**94.33**
	OPERA	96.59	77.47	92.89	77.47	88.03	77.47
	MASH_VLM	95.13	74.27	92.76	74.27	87.29	74.20
	SELF_CHECK	96.94	76.00	92.91	76.00	88.62	75.80
	SLIDING	**99.35**	60.73	**98.17**	60.73	**95.98**	60.47
	VIGOR	96.83	81.47	92.79	81.47	88.01	81.27
LLaVA-1.5-13B	Greedy	97.16	77.60	94.02	77.60	89.74	77.53
	VCD	90.94	89.67	83.07	89.93	78.04	89.80
	DECO	77.24	**95.27**	70.92	**95.27**	67.26	**95.20**
	OPERA	94.54	84.33	89.21	84.33	83.55	84.33
	MASH_VLM	94.85	78.60	91.89	78.60	87.08	78.67
	SELF_CHECK	97.46	76.87	95.13	76.87	91.20	76.73
	SLIDING	**98.49**	69.47	**97.11**	69.47	**94.64**	69.40
	VIGOR	96.00	84.87	92.05	84.87	86.37	84.93
Qwen-VL-Chat	GREEDY	98.49	74.07	95.36	74.07	91.60	74.13
	VCD	93.16	88.07	84.31	88.13	78.50	88.13
	DECO	89.22	**91.00**	84.94	**91.00**	75.44	**90.93**
	OPERA	97.26	78.13	93.39	78.13	87.91	78.07
	MASH_VLM	96.65	75.00	93.52	75.00	89.16	75.13
	SELF_CHECK	98.56	73.00	95.72	73.00	92.11	73.20
	SLIDING	**99.19**	65.20	**97.80**	65.20	**96.07**	65.27
	VIGOR	97.59	80.93	94.40	80.93	88.98	81.27
Qwen2.5-VL-7B	GREEDY	98.73	77.67	96.44	77.67	94.57	77.73
	VCD	98.19	79.67	96.14	79.80	94.04	79.93
	DECO	79.93	**97.73**	71.41	**97.73**	65.00	**97.67**
	OPERA	93.54	84.93	87.32	84.93	83.63	85.13
	MASH_VLM	97.42	80.40	94.44	80.40	91.50	80.40
	SELF_CHECK	98.99	78.07	96.62	78.07	94.89	78.00
	SLIDING	**99.80**	65.93	**98.02**	65.93	**97.34**	65.93
	VIGOR	98.25	82.47	95.15	82.47	92.10	82.40
Phi-3.5-vision	GREEDY	97.96	73.80	95.27	73.80	91.41	73.80
	VCD	91.95	85.33	86.94	85.20	82.64	85.07
	DECO	78.40	**92.93**	75.80	**92.93**	71.22	**92.87**
	OPERA	97.94	73.07	95.39	73.07	91.63	73.00
	MASH_VLM	95.80	74.60	93.95	74.60	90.61	74.60
	SELF_CHECK	98.43	71.27	96.22	71.27	92.39	71.27
	SLIDING	**98.92**	67.27	**97.68**	67.27	**95.09**	67.13
	VIGOR	96.00	81.67	93.37	81.67	88.52	81.73

cupies only a very small image region. Under this low-resolution condition, both the global view and the baseline grid-based crop produce false negative predictions. The baseline grid layout again places the target near a crop boundary, making the already weak evidence harder to preserve. By comparison, the VIGOR crop layout captures a more appropriate local region around the target and corrects the prediction to "YES" with a confidence score of 0.69.

Overall, this qualitative visualization shows that VIGOR can reduce the truncation and boundary-placement limitations of rigid grid-based cropping. By preserving tiny targets and their surrounding context within more suitable local views, VIGOR is more robust for detecting sparse visual evidence in challenging maritime scenes.

B Additional Analysis

B.1 Complexity Analysis

Let C_{base} denote the cost of one global pass, let C_{crop} denote the cost of processing one local crop, and let C_{fuse} denote the negligible cost of top-K selection, score fusion, stabilization, and fallback. Let

$$q = P(\mathbb{G}(v_g, \mathbf{X}) = 1) \tag{24}$$

be the routing rate induced by the sufficiency gate in Eq. 7. The expected cost of VIGOR is

$$\mathbb{E}[C] = C_{base} + q\,(MC_{crop} + C_{fuse})\,. \tag{25}$$

By contrast, an always-on multi-view baseline incurs

$$C_{all} = C_{base} + MC_{crop} + C_{fuse}. \tag{26}$$

The resulting relative speedup of VIGOR over

always-on refinement is

$$\text{Speedup} = \frac{C_{all}}{\mathbb{E}[C]}. \tag{27}$$

VIGOR therefore retains the benefit of multi-view refinement while paying the multi-view cost only on routed samples.

B.2 Mechanistic Localization

In this section, we provide the full methodological details of the mechanistic localization experiment, along with additional results on the most hallucination-prone setting (COCO-Adversarial) and discussion.

From a mechanistic interpretability perspective, hallucination and correct cognition are often simultaneously embedded in the residual stream. In this part, we construct the two poles of a binary response as two token subsets. We then design a layer-wise mechanistic localization analysis. Specifically, we extract the residual stream at each layer and project it into the logit space through Vocabulary Projection. For each of the two subsets, we sum the logits of all included tokens. We then compute the difference between the summation corresponding to the correct polarity and that corresponding to the hallucinated polarity, based on the ground-truth label. We report the averaged results over test samples of all five datasets (three COCO splits, MME, and GQA-Binary).

The results show that, compared with three competitive baseline methods, the intervention by VIGOR more effectively drives the model to gradually accumulate correctly recognized visual evidence and reasoning outcomes into the residual stream during propagation. We argue that this is because the representations injected by VIGOR's local views into the residual stream exert a calibration effect on the visually grounded concepts encoded in it. In addition, the results also indicate that VIGOR suppresses the forming of hallucination signals at an earlier stage. Part of this effect even emerges before reaching the layers for reasoning. This finding further explains why, in the previous evaluations, the performance advantage of VIGOR is amplified under hard settings that place stronger demands on visual reasoning. Furthermore, the separate results on COCO-Adversarial (shown in Figure 10) indicate that, as peripheral or ambiguous targets become more prevalent in the image, our method exerts a stronger corrective effect during inference than VCD, the strongest concept-disentanglement baseline, while not significantly changing the model's behavioral pattern.

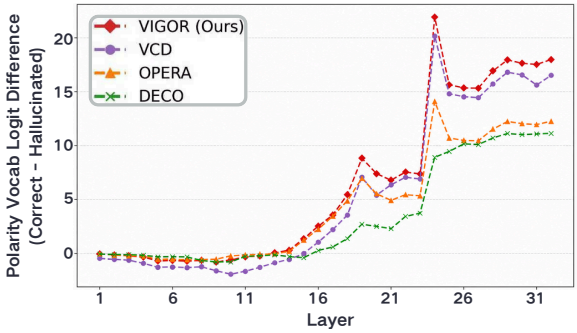


Figure 10: Layer-wise mechanistic localization analysis on POPE-COCO-Adversarial (Object: Residual stream; Base model: LLaVA-1.5-7B; Metric: Logit difference (correct - hallucinated)).

B.3 Probing Analysis (Early-Stage)

In addition to mechanistic localization, we further analyze intermediate residual representations with linear probes. Figure 11 shows that, compared with baselines built on the same LLaVA-1.5-13B backbone, VIGOR reduces the early linear decodability regarding object verification in layer 13-18, especially on the POPE-COCO Popular and Adversarial splits. This pattern is consistent with our hypothesis that selective local refinement delays premature commitment to weakly grounded object predictions. We also observe that the turning point in the probing curves roughly aligns with the stage at which the mechanistic localization analysis coincidentally begins to show a clearer positive separation between correct and hallucinated response poles. Taken together, these results suggest that VIGOR not only changes final predictions, but also alters earlier latent process in a way that is more patient and more consistent with visually grounded verification.

B.4 Error Analysis

To better understand the source of VIGOR's performance gains, we further conduct an error analysis comparing VIGOR with Greedy on the Adversarial split of POPE using LLaVA-13B. Tables 14 and 15 show that the performance improvement of VIGOR is not uniformly distributed across all error types, but is primarily driven by its ability to recover positive cases missed by Greedy. Compared with Greedy, VIGOR improves from 84.63 to 86.33, recall from 77.20 to 84.60, and F1-score from 83.40 to 86.09, while precision decreases only

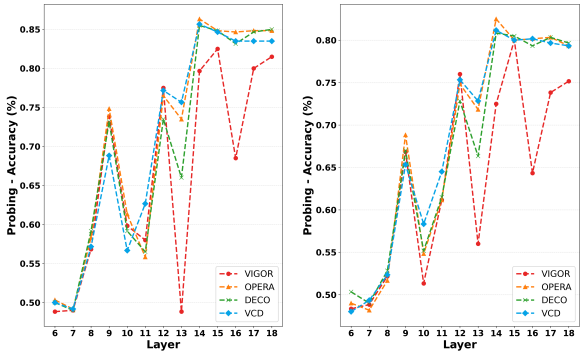


Figure 11: Layer-wise probing analysis of intermediate representations on the POPE-COCO Popular (left) and Adversarial (right) splits, using LLaVA-1.5-13B as the backbone.

moderately from 90.68 to 87.64. This pattern indicates that VIGOR shifts the prediction behavior toward greater detection sensitivity. The outcome breakdown in Table 15 further supports this interpretation: although VIGOR introduces 76 new errors, it corrects 127 errors made by Greedy. Notably, most of the corrected cases are concentrated in ground-truth positive samples, with 117 out of 127 fixed cases belonging to the positive class. This indicates that the main contribution of VIGOR lies in recovering positive evidence that is overlooked under the global view. Consistent with this observation, Table 14 shows a substantial increase in true positives (from 1158 to 1269) and a marked reduction in false negatives (from 342 to 231), whereas the increase in false positives (from 119 to 179) remains relatively limited. Overall, these results suggest that multi-view refinement improves performance primarily by reducing missed positives rather than by uniformly improving all prediction outcomes.

C Technical Details

C.1 Datasets

We evaluate VIGOR against several baselines on two widely used benchmarks for object hallucination: POPE (Li et al., 2023b) and MME (Fu et al., 2024).

C.1.1 POPE

POPE (Li et al., 2023b) evaluates object hallucination in the VQA setting by asking LVLMs whether specific objects are present, while maintaining a strict 1:1 ratio between existent and nonexistent objects. It includes three negative-sample sampling strategies—*Random*, *Popular*, and *Adversarial*—with increasing difficulty, thereby testing the model's robustness to statistical priors. In our experiments, we use the evaluation subsets constructed from MSCOCO (Lin et al., 2014) and GQA (Hudson and Manning, 2019).

C.1.2 MME

MME (Fu et al., 2024) is a comprehensive benchmark for evaluating LVLMs, covering 14 subtasks that assess both perceptual and cognitive abilities. We focus on five subsets that are most relevant to our motivations: (1) Existence and (2) Count, which evaluate basic visual robustness and the ability to disambiguate crowded objects through localized sparsity; (3) Position, which tests spatial reasoning, as our multi-view cropping provides explicit local coordinate references; (4) Posters, which examines the recovery of high-frequency, OCR-level details that are often weakened by global pooling; and (5) Scene, which serves as a key diagnostic subset for verifying that the global-local alignment in Stage III, through the fusion of w_g and w_l, preserves holistic semantic understanding without sacrificing global context. We report the data statistics in Table 16.

C.2 Baselines

In addition to the default **Greedy** decoding baseline, we compare VIGOR with six competitive training-free hallucination mitigation methods: **VCD** (Leng et al., 2024), **DeCo** (Wang et al., 2025a), **OPERA** (Huang et al., 2024a), **MASH-VLM** (Bae et al., 2025), **SELF_CHECK** (Manakul et al., 2023), and **SLIDING**. Among them, VCD and MASH-VLM are representative visual-contrastive methods that amplify grounded signals through perturbed visual inputs. OPERA and DeCo are recent methods that mitigate hallucinations through over-trust attention penalization and dynamic logit correction, respectively. In addition, we include SELF_CHECK and SLIDING as designed baseline variants to represent post-hoc verification and spatial multi-view fusion strategies. They serve as controlled training-free comparisons to test whether the gains of VIGOR arise simply from an additional verification pass or naive multi-view fusion, rather than from its selective routing and evidence integration design. Self_Check performs a standard forward pass to obtain the initial prediction and its confidence, measured by the yes–no logit margin. It then conducts a second inference with an added verification prompt (e.g.,

Table 14: Quantitative comparison between Greedy and VIGOR on the Adversarial split of POPE using LLaVA-13B.

Method	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	TP	TN	FP	FN
Greedy (NONE)	84.63	90.68	77.20	83.40	1158	1381	119	342
VIGOR (Full)	86.33	87.64	84.60	86.09	1269	1321	179	231

Table 15: Prediction outcomes of Greedy and VIGOR on the POPE Adversarial split using LLaVA-13B.

Category	Count	Predicted Yes	Predicted No
Both Correct	2463	1152	1311
Fixed by VIGOR	127	117	10
Broken by VIGOR	76	6	70
Both Wrong	334	225	109

"Please verify carefully.") to obtain a verification prediction and confidence. The final decision is made by comparing the two stages: consistent predictions are kept, while the verification result is adopted only when the initial confidence is very weak or when the verification confidence exceeds the initial one by a sufficient margin. Otherwise, the original prediction is retained. SLIDING combines global and local visual evidence through simple multi-view fusion. It first performs inference on the original image to obtain global logits, and then separately evaluates four non-overlapping quadrant crops. The logits from the four local crops are averaged to form a local-view estimate. The final prediction is produced by equally weighting the global logits and the averaged local logits.

C.3 Implementation Details.

VIGOR is training-free and does not update any parameters of the underlying MLLM. In our implementation, the global sufficiency gate uses a fixed trigger margin of $\delta = 3.6$, the local branch retains the top-K local probabilities, with $K = 2$, and the global-local fusion weights are set to $(w_g, w_l) = (0.52, 0.48)$. For asymmetric stabilization, we use $(\eta_s, \nu_s, \Delta_s) = (0.8, 3, 2.0)$ for strong positive recovery, $(\eta_m, \nu_m, \Delta_m) = (0.2, 4, 1.4)$ for medium recovery, $(\eta_l, \nu_l, \Delta_l) = (0.25, 2, 0.9)$ for light recovery, and $(\eta_n, \nu_n, \Delta_n) = (-1.0, 5, 1.0)$ for negative correction. The low-margin fallback uses $(\epsilon, \Delta_f) = (0.30, 0.25)$. Once initialized, VIGOR operates purely at inference time as a wrapper around the forward passes of the underlying model, without retraining.

For fair comparison, we adopt the standard or optimized settings of the baselines. DeCo uses a dynamic correction coefficient $\alpha = 0.6$, with candidate selection thresholds of $p = 0.9$ (Top-p) and $k = 20$ (Top-k). VCD uses a visual contrastive weight of $\alpha = 1.0$. MASH-VLM uses $\alpha = 1.0$ with a Gaussian blur radius of 2.0. OPERA uses an over-trust penalty coefficient of 1.0, a column-wise attention sum threshold of 25.0, and monitors the top 3 attention layers with a 2-step warmup. SELF_CHECK uses a low-confidence band of 0.50 and a verification margin of 0.10. The SLIDING multi-crop baseline uniformly balances global and average local predictions with fusion weights of 0.5.

D Literature Review

D.1 Hallucination in Large Language Models

Recent years have witnessed the unprecedented fluency and logical reasoning capabilities of Large Language Models in natural language generation tasks (Brown et al., 2020; Ouyang et al., 2022). However, this powerful generative ability is accompanied by the pervasive issue of "hallucination." In their comprehensive survey, Ji et al. (2023) define hallucination in natural language generation as the production of text that is nonsensical or unfaithful to the source content; later surveys further refine the taxonomy of hallucinations in the LLM era (Huang et al., 2025). However, it presents an illusion of fluency and naturalness. In text-only LLMs, hallucination often comes from factual errors in pre-training data and a lack of long-tail knowledge (Maynez et al., 2020; Li et al., 2023a; Huang et al., 2025). Because LLMs tend to predict the next token by maximizing statistical probability, they frequently rely on internal language priors, leading to plausible but wrong outputs when lacking contextual support. This issue roughly restricts the reliability for deployment of LLMs in high-risk domains (Ouyang et al., 2022; Bender et al., 2021).

D.2 Hallucination in MLLMs and Mitigation Efforts

With the integration of visual modules and LLMs, MLLMs have achieved remarkable cross-modal understanding and reasoning. Nevertheless, intro-

Dataset	Split / Subset	Description	Samples
POPE (COCO) (Li et al., 2023b)	Random	Objects sampled uniformly at random	3,000
	Popular	Frequent objects to test statistical bias	3,000
	Adversarial	Co-occurring objects acting as hard negatives	3,000
	Total	*Full POPE COCO evaluation set*	*9,000*
MME (Fu et al., 2024)	Yes/No Subset	Comprehensive perception subtasks (binary)	2,374
GQA-Binary (Hudson and Manning, 2019)	Validation Split	Yes/No + Binary semantic subset	24,783

Table 16: Statistics of the evaluation datasets used in our experiments. For POPE, we use the standard full configuration on the COCO split sourced from the lmms-lab benchmark. For MME, we focus on the complete yes/no subset. For GQA, we evaluate on the binary yes/no and semantic questions from the validation split.

ducing a new model pattern does not eliminate hallucinations. It may trigger more complex hallucinations in the vision-language due to the inherent challenges of Multimodal alignment (Liu et al., 2023; Chen et al., 2024; Huang et al., 2024b). In MLLMs, models may not only fabricate facts but also generate descriptions that contradict with the input image (Huang et al., 2024b; Chen et al., 2024). Previous research indicates that mitigating hallucinations from MLLM is a highly challenging task. During the training phase, researchers typically employ robust instruction tuning with high-quality positive and negative samples or Reinforcement Learning from Human Feedback (RLHF) to forcefully align image-text relationships (Liu et al., 2023; Gunjal et al., 2024; Sun et al., 2024; Xiao et al., 2025). During the inference phase, many studies focus on suppressing hallucinations by intervening in the decoding process without retraining the model (Leng et al., 2024; Huang et al., 2024a; Wang et al., 2024; Favero et al., 2024; Liu et al., 2024b). However, due to the inherent structural differences between visual encoding and textual decoding, purely global alignment often fails to eliminate the model's groundless assumptions regarding corresponding visual details (Favero et al., 2024; Liu et al., 2024b; Min et al., 2025).

D.3 Object Hallucination in MLLMs and Mitigation Strategies

Among the various expressions of hallucinations in MLLMs, "object hallucination" is one of the most stubborn and fatal failure modes (Rohrbach et al., 2018). Li et al. (Li et al., 2023b) conducted the first systematic study of this phenomenon and proposed the widely adopted POPE benchmark; subsequent benchmarks such as HallusionBench further broadened diagnostic coverage beyond simple object polling (Guan et al., 2024). Their research demonstrates that MLLMs are highly susceptible to visual instructions and object co-occurrence frequencies (Rohrbach et al., 2018). For instance, if a dining table is present in the image, the model is highly prone to hallucinating a chair based on statistical common sense, even if there is no chair existing. To alleviate object hallucinations, particularly in a training-free manner, a plethora of decoding intervention strategies have recently emerged (Leng et al., 2024; Huang et al., 2024a; Wang et al., 2024; Favero et al., 2024; Liu et al., 2024b; Zhong et al., 2024). For example, Leng et al. introduced Visual Contrastive Decoding (VCD) (Leng et al., 2024), which effectively reduces the model's over-reliance on statistical biases and unimodal priors by contrasting the output distributions of original and distorted images. However, such decoding-level interventions typically only offer post-hoc token-level corrections after the visual representations have been formed. Fundamentally, they cannot recover fine-grained visual evidence that was unsuccessfully encoded in the first place.

D.4 Causal Analysis and Diagnostic Localization of Object Hallucination

To fundamentally resolve object hallucinations, recent studies have begun to delve into the internal network architectures of MLLMs to perform "diagnostic localization" of where hallucinations originate (Yang et al., 2025; Jiang et al., 2025). Overall, the deep-seated mechanisms of object hallucination primarily center on two aspects: architectural resolution bottlenecks and progressive visual context dilution (Jiang et al., 2025; Li et al., 2025; Favero et al., 2024).

First, regarding the vision encoder, because MLLMs typically utilize fixed-resolution visual encoders, high-resolution images must be forcibly downsampled and compressed into a limited num-

ber of visual tokens. As argued by Guo et al. (2024b), this architectural bottleneck weakens the representation of subtle local details, meaning the lack of reliable visual evidence directly induces fine-grained grounding failures (Guo et al., 2024b). Second, from the perspective of the generation process, diagnostic studies reveal that visual attention is not completely static. As discovered by Li et al. (Li et al., 2025; Favero et al., 2024; Min et al., 2025), visual information is often progressively diluted during long-sequence generation. This means that even if the one-shot global visual encoding captures some initial information, it becomes inadequate for reliable grounding as the text sequence becomes longer. Furthermore, fine-grained distinctions that are not reliably captured by the vision encoder leave MLLMs with highly unstable grounding signals (Jiang et al., 2025; Yang et al., 2025). As a result, as visual evidence fades, the model is forced to fall back on language priors to complete the sequence, thereby hallucinations are triggered.

In light of these diagnostic findings, we argue that since the root cause of object hallucination lies in the lack of local visual evidence due to initial fixed-resolution constraints and progressive visual dilution during generation, relying solely on post-hoc global decoding penalties is not enough. This motivates our proposed VIGOR framework. By introducing active perception through an uncertainty-driven gating mechanism, VIGOR precisely intervenes when the initial global assessment becomes uncertain. It uses local crops to provide high-resolution, incremental visual evidence, thereby restoring fine-grained visual grounding without paying the cost of unnecessary computational overhead.

E Future Work

Based on the findings of our study, future work can proceed in several directions.

First, future work could extend VIGOR beyond constrained binary verification to more open-ended generation tasks, such as visual question answering and image captioning. In these settings, the model no longer produces a single yes-or-no decision, but instead generates responses token by token.

Second, future work could explore a dynamic intervention mechanism that monitors generation-time uncertainty and selectively triggers local refinement. Rather than relying on a fixed existential margin, the mechanism could be guided by sequence-level uncertainty signals, such as predictive entropy or fluctuations in top-token probabilities.

Third, future work could explore more adaptive cropping strategies to reduce the computational redundancy introduced by exhaustive grid-based crops. For example, attention-guided or saliency-driven patch selection may enable local refinement to focus only on regions associated with high uncertainty or dense semantic information.

Lastly, future work could also extend VIGOR from static images to spatio-temporal settings, where visual evidence dilution may become even more severe. One possible next step is to adapt its asymmetric existential aggregation to video inputs, so that local evidence can be integrated more effectively across frames. This may improve grounding in multi-frame MLLMs, where visual context accumulates quickly over time.