VAUQ: Vision-Aware Uncertainty Quantification for LVLM Self-Evaluation

Seongheon Park[1] **Changdae Oh**[1] **Hyeong Kyu Choi**[1]
Sean Du[2] **Sharon Li**[1]
[1]University of Wisconsin-Madison [2]Nanyang Technological University
{seongheon_park, sharonli}@cs.wisc.edu

Abstract

Large Vision-Language Models (LVLMs) frequently hallucinate, limiting their safe deployment in real-world applications. Existing LLM self-evaluation methods rely on a model's ability to estimate the correctness of its own outputs, which can improve deployment reliability; however, they depend heavily on language priors and are therefore ill-suited for evaluating vision-conditioned predictions. We propose **VAUQ**, a vision-aware uncertainty quantification framework for LVLM self-evaluation that explicitly measures how strongly a model's output depends on visual evidence. VAUQ introduces the Image-Information Score (IS), which captures the reduction in predictive uncertainty attributable to visual input, and an unsupervised core-region masking strategy that amplifies the influence of salient regions. Combining predictive entropy with this core-masked IS yields a training-free scoring function that reliably reflects answer correctness. Comprehensive experiments show that VAUQ consistently outperforms existing self-evaluation methods across multiple datasets[1].

1 Introduction

LVLMs have demonstrated remarkable progress across a wide range of multimodal tasks, exhibiting strong visual understanding capabilities. However, LVLMs remain prone to hallucinations, posing significant risks in high-stakes domains (Liu et al., 2024b). To assess model outputs, many existing works rely on external evaluators and judges (Liu et al., 2024a; Lee et al., 2024). Nevertheless, this approach is both costly and susceptible to hallucinations from the evaluator itself (Xu et al., 2024).

A promising direction for improving deployment reliability is *self-evaluation*, in which a model estimates the correctness of its own outputs using internal signals, without relying on external supervision or auxiliary models. In language-only settings,

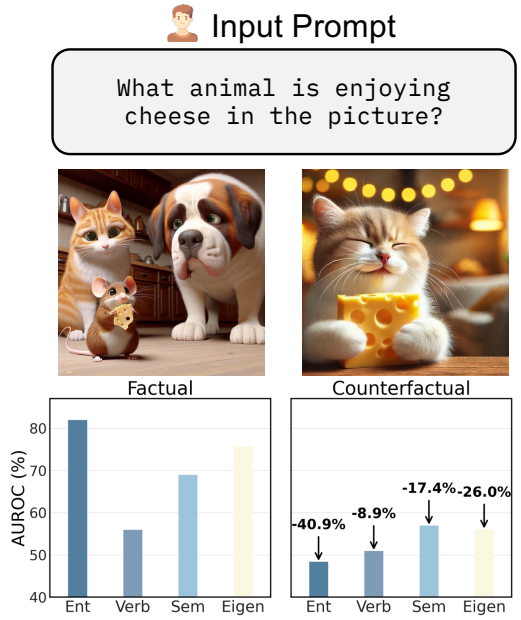


Figure 1: **Failure of LLM-based self-evaluation under language prior dominance.** Methods include: Entropy (Ent), Verbalized Confidence (Verb), Semantic Entropy (Sem), and EigenScore (Eigen). Performance comparison on the ViLP dataset using LLaVA-1.5-7B, which contains paired factual and counterfactual images associated with the same prompt. Common self-evaluation methods often fail in counterfactual samples.

prior work has shown that uncertainty quantification, consistency checks, and latent-state analysis can enable large language models to identify unreliable generations of themselves (Malinin and Gales, 2021; Manakul et al., 2023; Kuhn et al., 2023; Du et al., 2024; Orgad et al., 2025; Wang et al., 2025b). However, extending these self-evaluation techniques to LVLMs is non-trivial. Unlike text-only models, LVLMs operate over heterogeneous modalities, and their uncertainty arises from both linguistic and visual sources.

In practice, LVLMs can exhibit a strong language-prior dominance, relying heavily on statistical regularities learned during large-scale language pretraining while under-utilizing visual ev-

[1]https://github.com/deeplearning-wisc/vauq

Findings of the Association for Computational Linguistics: ACL 2026, pages 26534–26550

idence (Liu et al., 2024c; Leng et al., 2024; Lee et al., 2025; Long et al., 2025). As a result, existing LLM-based self-evaluation methods can assign low uncertainty to hallucinated responses when the image contradicts common linguistic expectations (see Figure 1). In such cases, confidence reflects fluency rather than grounding. This fundamental mismatch highlights the need for an approach that explicitly accounts for how much visual information contributes to a model's prediction.

To this end, we introduce **V**ision-**A**ware **U**ncertainty **Q**uantification (**VAUQ**), a training-free framework for LVLM self-evaluation that measures whether a model's confidence is justified by visual grounding. Our central insight is that informative and correctly utilized visual evidence should reduce predictive uncertainty on LVLM's output. VAUQ operationalizes this insight through two key components. First, we propose the Image-Information Score (IS), which quantifies the reduction in predictive uncertainty attributable to the image by comparing model behavior with and without visual input. Moreover, to ensure that IS reflects semantically meaningful visual information rather than spurious background correlations, we introduce an unsupervised core-region masking strategy. By selectively masking the salient regions, VAUQ penalizes predictions that remain confident even after core visual evidence is removed.

We evaluate VAUQ across diverse benchmark datasets and widely used LVLMs, including LLaVA (Liu et al., 2023), Qwen2.5-VL (Bai et al., 2025), and InternVL3.5 (Wang et al., 2025a). Across all settings, VAUQ consistently outperforms existing LLM- and LVLM-based self-evaluation methods. In the challenging counterfactual scenarios where visual grounding is essential, our approach achieves a significant **+13.3%** improvement in self-evaluation AUROC compared to state-of-the-art methods. We conduct comprehensive ablation studies that systematically analyze each component of VAUQ, demonstrating the necessity and robustness of each design choice. Our key contributions are summarized as follows:

1. We propose VAUQ, a novel vision-aware uncertainty quantification framework that enables LVLMs to perform reliable self-evaluation without relying on external models.

2. We introduce an information-theoretic score together with a core-region masking strategy to effectively capture visual utilization in a label-free and training-free manner.

3. We conduct extensive experiments across multiple LVLMs and benchmark datasets, achieving state-of-the-art performance and providing a rigorous analysis of the contributions of the proposed components.

2 Related Works

LLM self-evaluation aims to assess the correctness of a model's own outputs using only its internal knowledge, without relying on external supervision. This is closely related to uncertainty quantification and hallucination detection, which can be broadly categorized as follows: (1) logit-based methods, which use token-level probabilities as uncertainty scores (Malinin and Gales, 2021; Orgad et al., 2025); (2) prompting-based methods, which instruct models to explicitly express their uncertainty in natural language (Kadavath et al., 2022; Lin et al., 2022, 2024); (3) consistency-based methods, which evaluate uncertainty by measuring the agreement across multiple generated responses (Manakul et al., 2023; Kuhn et al., 2023; Chen et al., 2024; Li et al., 2025); and (4) internal-state-based methods, which leverage latent representations to assess hallucination (Ren et al., 2022; Azaria and Mitchell, 2023; Burns et al., 2023; Du et al., 2024; Park et al., 2025; Wang et al., 2025b).

LVLM self-evaluation *remains largely under-explored*, as it is further complicated by the uncertainty introduced through the integration of the image information. Prior studies have examined token-level probabilities (Zhou et al., 2024), visual attention weights (Jiang et al., 2025), or latent representation (Phukan et al., 2025; Yang et al., 2025; Park and Li, 2025; Duan et al., 2025) to detect hallucinations at the object level. However, these methods are limited to object tokens and cannot effectively assess response-level hallucination, which is more generally applicable across diverse vision-language tasks. A few recent works attempt to detect response-level hallucination through linear probing (Li et al., 2024) or semantic-invariant perturbation (Khan and Fu, 2024; Zhang et al., 2025), but they rely on large labeled datasets, require computationally expensive multiple sampling, or external natural language inference modules.

In this paper, we propose an efficient framework for LVLM self-evaluation that identifies hallucinations based on the Image-information Score, en-

abling scalable and generalizable response-level assessment in a *training-free* manner, *without relying on external supervision*.

3 Problem Setup

Notation. Given an input image, the vision encoder processes it into a set of patch-level visual tokens. These tokens are then projected into the language model's embedding space through the multimodal fusion module, resulting in a sequence of N visual embeddings: $\mathbf{v} = \{v_1, \ldots, v_N\} \in \mathbb{R}^{N \times d}$, where each v_i corresponds to a transformed visual token of dimension d. On the language side, the input text prompt is tokenized and embedded into a sequence of language embeddings: $\mathbf{t} = \{t_1, \ldots, t_L\} \in \mathbb{R}^{L \times d}$, where L is the prompt length. The projected visual tokens $\mathbf{v}$ and the textual embeddings $\mathbf{t}$ are concatenated and passed as the input sequence to the language model. The language model then generates a sequence of output tokens: $\mathbf{y} = \{y_1, \ldots, y_M\}$, where each $y_i \in \mathcal{V}$ is drawn from a vocabulary space and M is the output length.

In real-world deployment, LVLMs are often required not only to generate answers, but also to assess the reliability of answers in the absence of external supervision. Self-evaluation is thus crucial for selective prediction, hallucination detection, and downstream decision-making under uncertainty. We provide the formal definition below.

Definition 3.1 (**LVLM Self-Evaluator**). *Let* $\mathbf{x} = (\mathbf{v}, \mathbf{t})$ *denote the multimodal input to an LVLM. The goal of self-evaluation is to design a scoring function* $s : \mathcal{X} \times \mathcal{Y} \to [0, 1]$*, where* $s(\mathbf{x}, \mathbf{y})$ *measures the likelihood that the generated response* $\mathbf{y}$ *is hallucinated for the input* $\mathbf{x}$*. Based on this score, we define the binary self-evaluator:*

$$G(\mathbf{x}, \mathbf{y}) = \begin{cases} 1, & \text{if } s(\mathbf{x}, \mathbf{y}) \geq \tau \\ 0, & \text{otherwise}, \end{cases} \tag{1}$$

where $\tau \in [0, 1]$ *is a decision threshold. A value of* $G(\mathbf{x}, \mathbf{y}) = 1$ *indicates that the response is classified as hallucinated (i.e., incorrect), while* $G(\mathbf{x}, \mathbf{y}) = 0$ *denotes a correct response.*

4 Our Approach

In this section, we introduce Vision-Aware Uncertainty Quantification (**VAUQ**), a training-free self-evaluation framework that explicitly measures an LVLM's reliance on visual evidence when generating responses. We first present our motivation in Section 4.1, followed by a detailed walkthrough of the proposed method in Section 4.2.

4.1 Motivation: LLM-based Self-evaluation Methods Suffer from Language Prior

Recent works on uncertainty quantification or hallucination detection methods for LLMs demonstrate that models can self-evaluate the correctness of their responses through internal signals (Lin et al., 2022; Wang et al., 2025b). However, these methods are developed for text-only settings, and it remains unclear whether they can capture uncertainty that arises specifically from incorporating image information in vision-language tasks. To investigate the effectiveness of LLM-based self-evaluation metrics in vision-language multimodal setups, we conduct a pilot study (Figure 1) on the ViLP dataset (Luo et al., 2025), which contains pairs of factual images aligned with common knowledge and counterfactual images paired with identical prompts. Dataset details are provided in Appendix C.

We consider four representative LLM-based self-evaluation approaches: (1) Length-normalized Entropy (Ent) (Malinin and Gales, 2021), which measures the length-normalized entropy of the next-token distribution; (2) Verbalized Confidence (Verb) (Lin et al., 2022), obtained by parsing the model's self-reported numerical confidence; (3) Semantic Entropy (Sem) (Kuhn et al., 2023), defined as the entropy over semantic clusters formed from multiple sampled responses; and (4) EigenScore (Eigen) (Chen et al., 2024), which quantifies representation-level variability via dominant eigenvalues of the hidden-state covariance matrix. All the formulations are provided in Appendix E.1.

We observe that all four metrics show substantial performance degradation on counterfactual images, *e.g.*, Entropy decreases by **-40.9%** and EigenScore by **-26.0%**, suggesting that these methods are strongly driven by language priors and fail to reliably incorporate visual evidence when measuring uncertainty. In particular, low predictive uncertainty does not necessarily indicate correct visual grounding, as an LVLM may remain confident even when its response is weakly supported, or contradicted, by the given image. These findings motivate the need for new metrics that *explicitly capture the use of visual evidence in LVLM self-evaluation*. Accordingly, our research question is:

How can we effectively quantify an LVLM's reliance on visual evidence for self-evaluation?

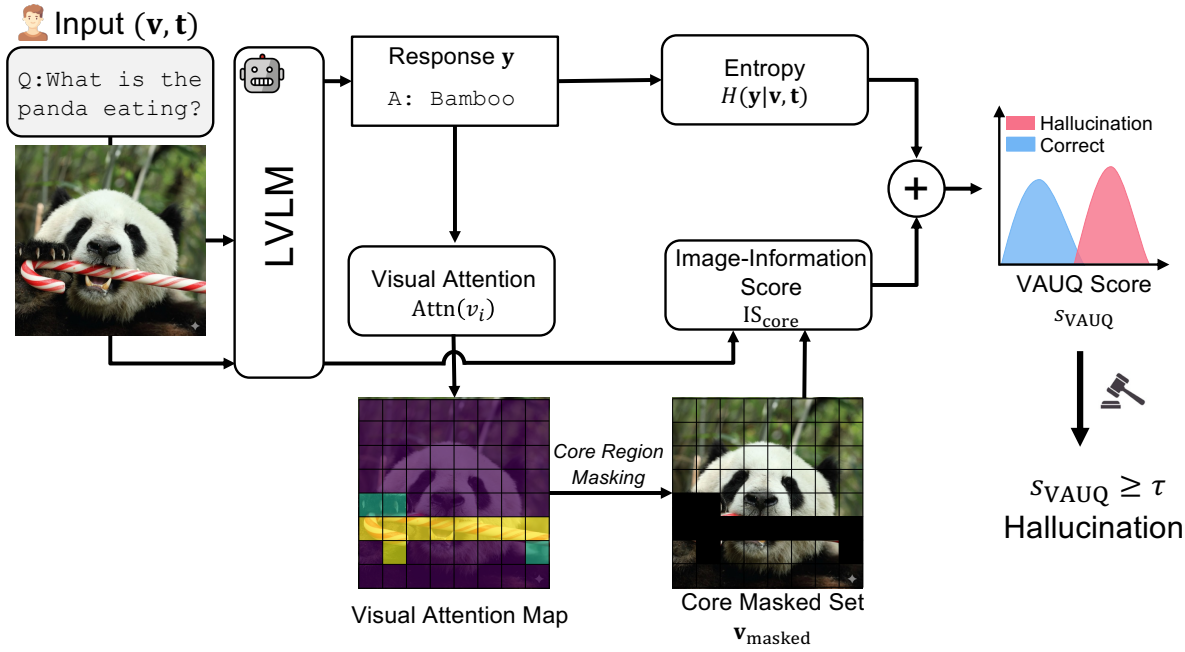


Figure 2: **Overall VAUQ Framework.** Given an input image-text pair $(\mathbf{v}, \mathbf{t})$, the LVLM generates a response $\mathbf{y}$. Based on the attention map $\mathrm{Attn}(v_i)$, we perform unsupervised core region masking by covering the top-$K\%$ image patches, resulting in a core-masked set $\mathbf{v}_{\text{masked}}$. Using this masked input, we compute the core-masked Image-Information Score $\mathrm{IS}_{\text{core}}$. Finally, predictive entropy $H(\mathbf{y} \mid \mathbf{v}, \mathbf{t})$ and $\mathrm{IS}_{\text{core}}$ are combined to produce the VAUQ score s_{VAUQ} for self-evaluation.

4.2 Vision-Aware Uncertainty Quantification

To address the research question, we propose a vision-aware uncertainty quantification framework that explicitly measures how much visual information contributes to a model's prediction. Our approach builds on the idea that visual evidence should reduce predictive uncertainty when it is informative and correctly utilized. Accordingly, we propose the *Image-Information Score* (IS), capturing the degree to which the image influences the model's predictive uncertainty.

Definition 4.1 (Image-Information Score (IS)). *Let $H(\mathbf{y} \mid \mathbf{v}, \mathbf{t})$[2] denote the length-normalized conditional entropy of the model's predictive distribution given visual features $\mathbf{v}$ and text $\mathbf{t}$, and let $H(\mathbf{y} \mid \varnothing, \mathbf{t})$ denote the entropy when visual tokens are removed. The Image-Information Score is defined as:*

$$\mathrm{IS}_{blank} = H(\mathbf{y} \mid \varnothing, \mathbf{t}) - H(\mathbf{y} \mid \mathbf{v}, \mathbf{t}), \quad (2)$$

where a larger IS indicates that visual information substantially reduces predictive uncertainty, reflecting stronger visual grounding during the models' answer generation.

[2] $H(\mathbf{y} \mid \mathbf{v}, \mathbf{t}) = -\frac{1}{M}\sum_{i=1}^{M}\sum_{y\in\mathcal{V}} p(y_i = y \mid \mathbf{y}_{<i}, \mathbf{v}, \mathbf{t}) \log p(y_i = y \mid \mathbf{y}_{<i}, \mathbf{v}, \mathbf{t})$

Unsupervised Core Region Masking. While $\mathrm{IS}_{\text{blank}}$ serves as an effective proxy for measuring image utilization, one limitation is that it can be sensitive to spurious correlations in the input (*e.g.*, background artifacts) (Yang et al., 2023), which introduce noise unrelated to the core visual evidence (See Table 3 for the details).

To mitigate this challenge, we introduce an unsupervised core-region masking strategy that focuses on the most task-relevant visual regions for score computation. Intuitively, if a model truly relies on the visual evidence needed to answer a query, removing that evidence should significantly increase predictive uncertainty. Since ground-truth evidence annotations are unavailable at inference time, we estimate these regions using the model's visual attention weights. Prior analyses indicate that middle to later transformer layers provide the most informative alignment between visual tokens and semantic reasoning (Jiang et al., 2025; Park and Li, 2025). We therefore aggregate attention from a contiguous range of layers, and empirically verify the optimality of this design choice in Section 5.3 (Figure 3).

Specifically, we define the interaction between a generated token y_i and visual information by summing the attention weights assigned to image to-

kens within an attention head h and layer ℓ:

$$\mathrm{Attn}(v_i) \triangleq \sum_{l=l_s}^{l_e} \sum_{h=1}^{H} \sum_{j=1}^{M} A^{(\ell,h)}(y_j, v_i), \quad (3)$$

where $A^{(\ell,h)}(y_j, v_i)$ denotes the attention weight from generated token y_j to image token v_i at the h-th attention head of the ℓ-th layer, M is the number of generated tokens, H is the total number of attention heads, and l_s and l_e indicate the start and end indices of the transformer layers used for aggregation, respectively.

We then select the top $K\%$ of image patches with the highest attention scores as $\mathbf{v}_{\text{top}}$:

$$\mathbf{v}_{\text{top}} \triangleq \left\{ v_i \;\middle|\; i \in \mathrm{TopK}\%\left(\{\mathrm{Attn}(v_i)\}_{i=1}^{N}\right) \right\}, \quad (4)$$

and define the *remaining* set of visual tokens, $\mathbf{v}_{\text{masked}}$, which is then used as input for core-masked IS computation:

$$\mathrm{IS}_{\text{core}} = H(\mathbf{y} \mid \mathbf{v}_{\text{masked}}, \mathbf{t}) - H(\mathbf{y} \mid \mathbf{v}, \mathbf{t}), \quad (5)$$

where $\mathbf{v}_{\text{masked}} \triangleq \{v_i \mid v_i \notin \mathbf{v}_{\text{top}}\}$. By masking the most attended regions, we intentionally remove the visual evidence the model is most likely to rely on, enabling a more precise assessment of whether its predictions are genuinely grounded.

Vision-Aware Uncertainty Quantification. Finally, we define the vision-aware uncertainty quantification score, s_{VAUQ}, as a linear combination of the predictive entropy and the IS score with the core region masking:

$$\begin{aligned} s_{\text{VAUQ}}(\mathbf{x}, \mathbf{y}) &= H(\mathbf{y} \mid \mathbf{v}, \mathbf{t}) - \alpha \cdot \mathrm{IS}_{\text{core}} \\ &= \underbrace{(1+\alpha)\, H(\mathbf{y} \mid \mathbf{v}, \mathbf{t})}_{\text{predictive uncertainty}} - \underbrace{\alpha\, H(\mathbf{y} \mid \mathbf{v}_{\text{masked}}, \mathbf{t})}_{\substack{\text{penalize weak use of} \\ \text{core visual information}}}, \end{aligned} \quad (6)$$

where α is a weighting hyperparameter. This formulation can be interpreted as measuring predictive uncertainty while explicitly discounting confidence that is not supported by core visual evidence. When a model relies on visual information, masking the core regions leads to a significant increase in entropy, resulting in a lower s_{VAUQ} score and signaling a more reliable, well-grounded prediction. In contrast, when a model's prediction is dominated by language priors, the IS score remains low—reflecting limited uncertainty reduction from visual input—which results in a relatively higher s_{VAUQ} value and indicates an increased risk of hallucination.

5 Experiments

5.1 Experimental Setup

Datasets and Models. We evaluate our method on three free-form visual question answering datasets: ViLP (Luo et al., 2025), MMVet (Yu et al., 2024), and VisualCoT (Shao et al., 2024), and one multiple-choice benchmark: CVBench (Tong et al., 2024). These datasets broadly cover the core challenges faced by deployed LVLMs, spanning language-prior dominance, multi-capability reasoning, evidence localization, and vision-centric perceptual reasoning.

We conduct experiments on three representative LVLMs: LLaVA-1.5-{7B, 13B} (Liu et al., 2023), Qwen-2.5-VL-7B (Bai et al., 2025), and InternVL3.5-8B (Wang et al., 2025a). Implementation details are provided in Appendix A.

Baselines. We compare our method against eight representative self-evaluation baselines spanning both LLM- and LVLM-based approaches. The LLM-based methods include Perplexity (Ren et al., 2022), Verbalized Confidence (Kadavath et al., 2022), Chain-of-Embeddings (Wang et al., 2025b), EigenScore (Chen et al., 2024), and Semantic Entropy (Kuhn et al., 2023). LVLM-based methods include SVAR (Jiang et al., 2025), Contextual Lens (Phukan et al., 2025), and VL-Uncertainty (Zhang et al., 2025).

Evaluation Metrics. Following prior work (Du et al., 2024; Park et al., 2025), we evaluate performance using the area under the receiver operating characteristic curve (AUROC). Ground-truth correctness labels for model responses are annotated using GPT-5 (OpenAI, 2025).

5.2 Main Experiments

VAUQ achieves state-of-the-art performance. In Tables 1 and 2, we compare VAUQ with competitive self-evaluation approaches from prior work. VAUQ achieves state-of-the-art performance, substantially outperforming existing methods on LLaVA, Qwen, and InternVL models. We observe that both LLM-based scoring approaches and object-level hallucination detectors (*e.g.*, SVAR and Contextual Lens) often exhibit inconsistent performance across architectures and data distributions. In contrast, VAUQ more consistently delivers strong results across model families, scales, and four benchmark datasets.

Method	LLaVA-1.5-7B				LLaVA-1.5-13B			
	ViLP	MMVet	VisualCoT	CVBench	ViLP	MMVet	VisualCoT	CVBench
Perplexity	54.6	79.3	56.2	60.3	54.2	81.3	63.7	64.6
Verbalized	56.3	71.7	49.9	57.6	52.4	67.2	54.4	55.4
SVAR	50.6	29.3	44.5	48.7	48.3	34.2	44.5	49.2
Contextual Lens	56.7	70.8	58.3	51.1	60.4	61.4	52.5	58.6
Chain-of-Embeddings	52.0	60.8	55.7	50.6	60.5	44.4	46.5	53.3
EigenScore	63.2	78.2	74.7	65.8	61.1	85.1	75.5	58.1
Semantic Entropy	63.7	81.3	75.1	70.2	63.5	86.4	75.7	66.2
VL-Uncertainty	55.6	**82.3**	65.2	71.1	58.6	85.9	77.7	67.9
VAUQ (Ours)	**77.0**	81.5	**77.8**	**73.2**	**69.5**	**88.6**	**80.2**	**68.3**

Table 1: **Main results with LLaVA.** Comparison with competitive self-evaluation methods across datasets. All values are AUROC (%). Best results are in **bold**, and second-best results are underlined.

Method	Qwen-2.5-VL-7B				InternVL3.5-8B			
	ViLP	MMVet	VisualCoT	CVBench	ViLP	MMVet	VisualCoT	CVBench
Perplexity	55.0	76.6	56.0	64.8	55.3	67.5	62.3	66.4
Verbalized	55.3	51.9	54.7	56.3	48.7	59.4	54.7	60.1
SVAR	49.6	54.6	46.7	61.6	51.6	49.7	56.7	50.5
Contextual Lens	61.3	65.5	65.0	54.4	51.3	59.6	55.1	48.2
Chain-of-Embeddings	47.4	59.8	44.3	49.7	50.1	52.8	61.4	53.6
EigenScore	53.0	60.8	51.1	50.9	59.3	64.1	74.5	61.0
Semantic Entropy	52.0	60.1	53.3	50.9	64.2	70.4	66.3	73.7
VL-Uncertainty	57.9	69.7	62.3	69.7	**67.4**	75.7	65.8	72.0
VAUQ (Ours)	**64.1**	**78.3**	**68.0**	**69.8**	65.2	**75.8**	**77.2**	**74.7**

Table 2: **Main results using QwenVL-2.5-7B and InternVL-3.5-8B.** All values are AUROC (%). Best results are in **bold**, and second-best results are underlined.

Notably, VAUQ improves over a representative LLM-based method Semantic Entropy by **+13.4%** on ViLP with LLaVA-1.5-7B. This highlights the importance of explicitly modeling visual evidence for uncertainty quantification for LVLMs. By jointly capturing predictive uncertainty and image-information utilization, VAUQ offers a more reliable and interpretable self-evaluation signal than prior text-centric methods. Moreover, VAUQ outperforms the previous state-of-the-art LVLM self-evaluation method, VL-Uncertainty, by **+21.4%** on ViLP and **+12.6%** on VisualCoT. While VL-Uncertainty estimates uncertainty by measuring semantic consistency across multiple sampled outputs and relies on external modules, VAUQ instead leverages internal model signals, requiring neither multiple sampling nor external components.

5.3 Ablation Studies

How effective is core-region masking? To investigate how masking strategies affect performance, we use the VisualCoT dataset (Shao et al., 2024), which provides ground-truth bounding boxes for the visual evidence required to answer each query. We consider three variants of masking: (1) *Blank*, which replaces the entire image with a blank input; (2) *Random*, which applies a random mask to the image; and (3) *GT (Oracle)*, which masks out the pixels inside the ground-truth evidence region.

Method	LLaVA-1.5	QwenVL-2.5	InternVL3.5
w. IS_{blank}	75.2	66.9	75.9
w. IS_{rand}	73.3	65.1	74.0
w. IS_{GT}	78.6	69.2	77.7
w. $\mathbf{IS_{core}}$	77.8	68.0	77.2

Table 3: AUROC (%) of VAUQ score under different masking strategies on the VisualCoT dataset.

The results in Table 3 show that masking the ground-truth evidence region (IS_{GT}) yields higher AUROC than the blank-image baseline (IS_{blank}), indicating that fine-grained and semantically relevant visual content is crucial for VAUQ score computation. In contrast, random masking (IS_{rand}) leads to degraded performance relative to the blank baseline, suggesting that indiscriminate perturbations disrupt the model's ability to extract meaningful visual cues. These findings highlight the importance of carefully selecting which visual regions to mask when estimating image-information utilization. Our method (IS_{core}) achieves performance comparable to the GT (Oracle) setting, demonstrating that our approximated masking region effectively captures the core visual evidence.

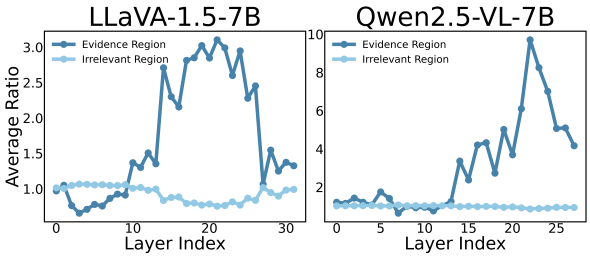


Figure 3: Visual attention ratios over evidence and irrelevant regions on the VisualCoT dataset.

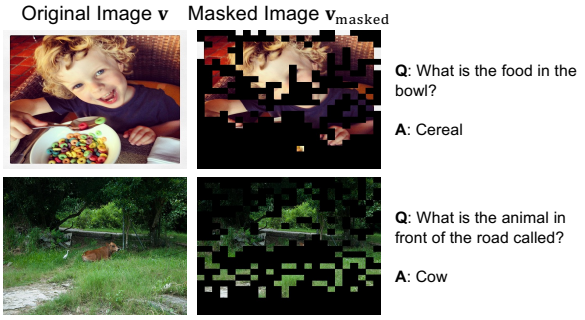


Figure 4: Qualitative examples of core region masking using LLaVA-1.5-7B.

Can visual attention weight capture the correct region? We investigate whether visual attention weights can capture the correct evidence region in the input image. Using the ground-truth evidence regions provided in the VisualCoT dataset, we compute the summed visual attention weight within the evidence region and within the irrelevant region, and then normalize each by the number of patches in the corresponding region. In Figure 3, we visualize the normalized attention weights for LLaVA-1.5-7B and Qwen2.5-VL-7B on VisualCoT. For both models, the early layers (0–10) struggle to focus on the correct region, while middle to later layers capture the evidence region more reliably. This observation is consistent with prior findings that intermediate layers of LVLMs are primarily responsible for processing visual information, and it further supports our design choice to use intermediate layers in our approach.

Visualization results of core region masking. We visualize masked images produced by the core region masking strategy using visual attention aggregated from the 10th-25th layers of LLaVA-1.5-7B, with $K = 60$. As shown in Figure 4, the masking strategy defined in Equation (4) effectively masks out semantically important regions required for reasoning, such as the cereal or the cow in the original image. These examples demonstrate that leveraging intermediate-layer visual attention can reliably identify meaningful visual regions. Additional qualitative results are provided in Appendix B.

Method	Factual		Counterfactual	
	LLaVA	QwenVL	LLaVA	QwenVL
Entropy	83.2	68.4	48.4	44.5
IS_{core}	60.9	56.5	75.2	63.0
Entropy $-\ \alpha\text{IS}_{\text{core}}$	83.6	68.1	70.4	60.1

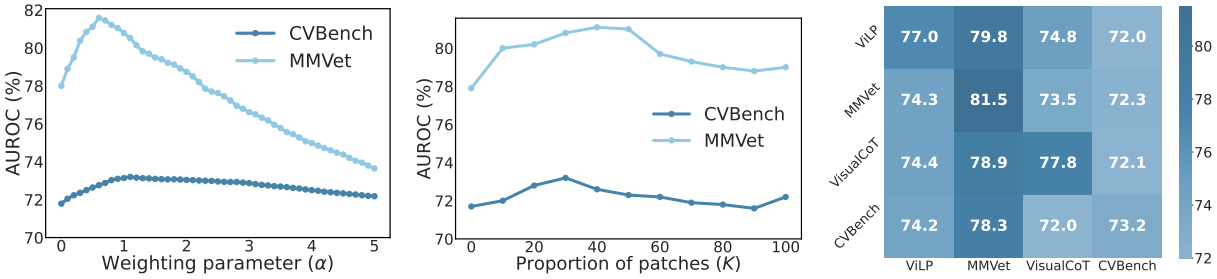
Table 4: Component analysis of VAUQ on the ViLP dataset across factual and counterfactual splits.

Method	LLaVA-1.5-7B		Qwen2.5-VL-7B	
	Time (s)	AUC (%)	Time (s)	AUC (%)
SVAR	0.39	50.6	1.59	49.6
Verbalized	0.58	56.3	1.82	55.3
EigenScore	5.86	63.2	8.77	53.0
Semantic Entropy	7.05	63.7	12.4	52.0
VL-Uncertainty	13.6	55.6	20.2	57.9
Ours	0.73	77.0	2.16	64.1

Table 5: Average per-sample inference time on the ViLP dataset.

Complementary roles of entropy and image information score. In Table 4, we conduct a component analysis of VAUQ on the ViLP dataset using LLaVA-1.5-7B and Qwen2.5-VL-7B. Entropy performs well on the factual split but degrades substantially on the counterfactual split due to its reliance on language priors. In contrast, IS_{core} achieves moderate performance on factual samples while exhibiting strong performance on the counterfactual split, where visual evidence is essential. By combining these two signals, VAUQ achieves robust and balanced performance across both factual and counterfactual settings. These results demonstrate that predictive entropy and image-information utilization capture complementary aspects of uncertainty, making VAUQ well-suited for mixed real-world distributions.

Inference efficiency of VAUQ. In Table 5, we evaluate the inference efficiency of VAUQ on the ViLP dataset. Generating a response of length M requires M forward passes under standard autoregressive decoding. VAUQ introduces only a constant number of additional forward passes for uncertainty estimation (e.g., computing scores with masked visual inputs), without requiring any additional autoregressive generation. Consequently, the overall inference complexity of VAUQ remains linear in the output length, $O(M)$, with a small constant overhead. In contrast, multi-sampling–based approaches require generating A indepen-



(a) Weighting parameter α (b) Proportion of patches K (c) Generalization results

Figure 5: (a) Effect of the weighting parameter α in Equation (6); (b) effect of the proportion of masked image patches K in Equation (4); (c) generalization performance across datasets.

dent responses, resulting in $O(A \cdot M)$ forward passes, where A is the number of samples and often exceeds five in practice (Zhang et al., 2025). Consequently, VAUQ achieves a 94.6% reduction in per-sample inference time compared to VL-Uncertainty, while delivering a +21.4% AUROC improvement in self-evaluation performance on the LLaVA-1.5-7B model. Overall, VAUQ maintains comparable per-sample inference time while providing stronger self-evaluation accuracy, demonstrating a favorable efficiency-performance trade-off across models.

How does the weighting parameter α affect performance? In Figure 5a, we present the performance across different values of the weighting parameter α. We search α over the range $[0, 5]$ with intervals of 0.1. Larger values of α place greater emphasis on the IS score relative to predictive entropy. In practice, moderate values of α between 0.5 and 1.5 yield the best performance, suggesting that predictive entropy and IS provide complementary and balanced contributions. Tasks that rely more heavily on visual evidence tend to prefer larger α values (*e.g.*, CVBench), while datasets requiring less visual grounding benefit from smaller values.

How does the proportion of the masked image patches K affect performance? We examine how the proportion of selected image patches $K\%$ used for masking influences model performance (see Figure 5b). We search K over the range [0, 100] with intervals of 10. We vary K from 0 to 100 in increments of 10. Overall, moderate values of K yield stable performance across datasets, with optimal performance achieved at $K = 30$ on CVBench and $K = 40$ on MMVet. In contrast, small masking regions tend to degrade performance, since they concentrate on overly specific areas or introduce noisy masks.

Parameter generalization across data distributions. While VAUQ demonstrates strong overall performance, we further investigate its ability to generalize across different data distributions. As shown in Figure 5c, we evaluate the generalization capability of VAUQ using the LLaVA-1.5-7B model by transferring hyperparameters—including the weighting parameter α, the proportion of masked image patches K—from a source in-distribution (ID) dataset to various target out-of-distribution (OOD) datasets, and computing their corresponding VAUQ scores. The results show that VAUQ transfers robustly across diverse datasets. Notably, it achieves an AUROC of 72.3% on CVBench even when the hyperparameters are selected from MMVet, closely matching the performance obtained when tuned directly on CVBench (73.2%). This transferability underscores VAUQ's practical potential for real-world LVLM applications, enabling effective self-evaluation even under significant domain shifts.

6 Conclusion

We introduced **VAUQ**, a vision-aware uncertainty quantification framework for LVLM self-evaluation that explicitly accounts for how much a model relies on visual evidence when assessing the reliability of its own outputs. Central to our approach is the Image-Information Score (IS), which measures the reduction in predictive uncertainty attributable to the image, and a core-region masking strategy that suppresses the influence of visually irrelevant regions during IS computation. Extensive experiments across multiple LVLM architectures and benchmark datasets show that VAUQ consistently outperforms existing LLM- and LVLM-based self-evaluation methods. We hope our work will encourage future research on reliable and vision-aware self-evaluation in multimodal models.

Limitations

Although VAUQ demonstrates robust performance across multiple LVLMs and benchmark datasets, it relies on a small set of global hyperparameters. While we show that VAUQ is relatively stable under moderate variations of these values, the optimal configuration can still vary across datasets and even individual examples, depending on how strongly a task relies on visual versus linguistic information. Hence, designing adaptive or sample-specific strategies for tuning these hyperparameters remains an interesting direction for future work.

Ethical Considerations

VAUQ aims to support more reliable use of vision–language models by providing a lightweight, training-free self-evaluation signal. Such a signal may be useful for identifying potentially unreliable outputs and for supporting selective prediction or human review in practical deployments. At the same time, VAUQ is not a comprehensive safety mechanism and should not be treated as a definitive measure of correctness. We view it as a complementary tool that is best used alongside existing safeguards and human oversight, particularly in sensitive or high-stakes settings.

Acknowledgements

We gratefully acknowledge Lin Long, Yu Wang, Wendi Li, and Dahye Kim for their valuable comments on the draft, and the anonymous reviewers for their constructive feedback. Seongheon Park, Changdae Oh, Hyeong Kyu Choi, and Sharon Li are supported in part by the AFOSR Young Investigator Program under award number FA9550-23-1-0184, National Science Foundation under awards IIS-2237037 and IIS-2331669, Office of Naval Research under grant number N00014-23-1-2643, Schmidt Sciences Foundation, Open Philanthropy, Alfred P. Sloan Fellowship, and gifts from Google and Amazon.

References

Wenbin An, Feng Tian, Sicong Leng, Jiahao Nie, Haonan Lin, QianYing Wang, Ping Chen, Xiaoqin Zhang, and Shijian Lu. 2025. Mitigating object hallucinations in large vision-language models with assembly of global and local attention. In *CVPR*.

Amos Azaria and Tom Mitchell. 2023. The internal state of an llm knows when it's lying. In *EMNLP Findings*.

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*.

Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. 2023. Discovering latent knowledge in language models without supervision. In *ICLR*.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: Llms' internal states retain the power of hallucination detection. In *ICLR*.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. 2023. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*.

Xuefeng Du, Chaowei Xiao, and Yixuan Li. 2024. Haloscope: Harnessing unlabeled llm generations for hallucination detection. In *NeurIPS*.

Jinhao Duan, Fei Kong, Hao Cheng, James Diffenderfer, Bhavya Kailkhura, Lichao Sun, Xiaofeng Zhu, Xiaoshuang Shi, and Kaidi Xu. 2025. Truthprint: Mitigating large vision-language models object hallucination via latent truthful-guided pre-intervention. In *ICCV*.

Zane Durante, Qiuyuan Huang, Naoki Wake, Ran Gong, Jae Sung Park, Bidipta Sarkar, Rohan Taori, Yusuke Noda, Demetri Terzopoulos, Yejin Choi, and 1 others. 2024. Agent ai: Surveying the horizons of multimodal interaction. *arXiv preprint arXiv:2401.03568*.

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 *CVPR*.

Shanghua Gao, Zhong-Yu Li, Ming-Hsuan Yang, Ming-Ming Cheng, Junwei Han, and Philip Torr. 2022. Large-scale unsupervised semantic segmentation.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *EMNLP*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

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 *CVPR*.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv e-prints*.

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 *CVPR*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Omri Kaduri, Shai Bagon, and Tali Dekel. 2025. What's in the image? a deep-dive into the vision of vision language models. In *CVPR*.

Zaid Khan and Yun Fu. 2024. Consistency and uncertainty: Identifying unreliable responses from black-box vision-language models for selective visual question answering. In *CVPR*.

Moo Jin Kim, Karl Pertsch, Siddharth Karamcheti, Ted Xiao, Ashwin Balakrishna, Suraj Nair, Rafael Rafailov, Ethan Foster, Grace Lam, Pannag Sanketi, and 1 others. 2024. Openvla: An open-source vision-language-action model. In *CoRL*.

Kimi Team. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *ICLR*.

Kang-il Lee, Minbeom Kim, Seunghyun Yoon, Minsung Kim, Dongryeol Lee, Hyukhun Koh, and Kyomin Jung. 2025. Vlind-bench: Measuring language priors in large vision-language models. In *NACCL Findings*.

Seongyun Lee, Seungone Kim, Sue Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. In *ACL Findings*.

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 *CVPR*.

Qing Li, Jiahui Geng, Chenyang Lyu, Derui Zhu, Maxim Panov, and Fakhri Karray. 2024. Reference-free hallucination detection for large vision-language models. In *EMNLP Findings*.

Xiaomin Li, Zhou Yu, Ziji Zhang, Yingying Zhuang, Swair Shah, Narayanan Sadagopan, and Anurag Beniwal. 2025. Semantic volume: Quantifying and detecting both external and internal uncertainty in llms. *arXiv preprint arXiv:2502.21239*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. In *TMLR*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. In *TMLR*.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *ICLR*.

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

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *NeurIPS*.

Sheng Liu, Haotian Ye, and James Zou. 2025. Reducing hallucinations in large vision-language models via latent space steering. In *ICLR*.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024c. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *ECCV*.

Xihui Liu, Zihao Wang, Jing Shao, Xiaogang Wang, and Hongsheng Li. 2019. Improving referring expression grounding with cross-modal attention-guided erasing. In *CVPR*.

Lin Long, Changdae Oh, Seongheon Park, and Sharon Li. 2025. Understanding language prior of lvlms by contrasting chain-of-embedding. *arXiv preprint arXiv:2509.23050*.

Yifan Lu, Ziqi Zhang, Chunfeng Yuan, Jun Gao, Congxuan Zhang, Xiaojuan Qi, Bing Li, and Weiming Hu. 2025. Mitigating hallucinations in large vision-language models by self-injecting hallucinations. In *EMNLP Findings*.

Tiange Luo, Ang Cao, Gunhee Lee, Justin Johnson, and Honglak Lee. 2025. Probing visual language priors in vlms. In *ICML*.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *ICLR*.

Potsawee Manakul, Adian Liusie, and Mark JF Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *EMNLP*.

OpenAI. 2025. Gpt-5: Large multimodal model by openai. https://openai.com/research/gpt-5.

Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. 2025. Llms know more than they show: On the intrinsic representation of llm hallucinations. In *ICLR*.

Seongheon Park, Xuefeng Du, Min-Hsuan Yeh, Haobo Wang, and Yixuan Li. 2025. Steer llm latents for hallucination detection. In *ICML*.

Seongheon Park and Yixuan Li. 2025. Glsim: Detecting object hallucinations in lvlms via global-local similarity. In *NeurIPS*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. In *NeurIPS*.

Anirudh Phukan, Harshit Kumar Morj, Apoorv Saxena, Koustava Goswami, and 1 others. 2025. Beyond logit lens: Contextual embeddings for robust hallucination detection & grounding in vlms. In *NACCL*.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *ICML*.

Jie Ren, Jiaming Luo, Yao Zhao, Kundan Krishna, Mohammad Saleh, Balaji Lakshminarayanan, and Peter J Liu. 2022. Out-of-distribution detection and selective generation for conditional language models. In *ICLR*.

Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. 2020. Grad-cam: visual explanations from deep networks via gradient-based localization. In *IJCV*.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multimodal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. In *NeurIPS*.

Yuge Shi, N Siddharth, Philip Torr, and Adam R Kosiorek. 2022. Adversarial masking for self-supervised learning. In *ICML*.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2024. Aligning large multimodal models with factually augmented RLHF. In *ACL Findings*.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, and 1 others. 2024. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. In *NeurIPS*.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025a. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Yiming Wang, Pei Zhang, Baosong Yang, Derek F Wong, and Rui Wang. 2025b. Latent space chain-of-embedding enables output-free llm self-evaluation. In *ICLR*.

Ziwei Xu, Sanjay Jain, and Mohan Kankanhalli. 2024. Hallucination is inevitable: An innate limitation of large language models. *arXiv preprint arXiv:2401.11817*.

Le Yang, Ziwei Zheng, Boxu Chen, Zhengyu Zhao, Chenhao Lin, and Chao Shen. 2025. Nullu: Mitigating object hallucinations in large vision-language models via halluspace projection. In *CVPR*.

Yu Yang, Besmira Nushi, Hamid Palangi, and Baharan Mirzasoleiman. 2023. Mitigating spurious correlations in multi-modal models during fine-tuning. In *ICML*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *ICML*.

Ruiyang Zhang, Hu Zhang, and Zhedong Zheng. 2025. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. In *CVPR*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. In *NeurIPS*.

Bolei Zhou, Aditya Khosla, Agata Lapedriza, Aude Oliva, and Antonio Torralba. 2016. Learning deep features for discriminative localization. In *ICCV*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *ICLR*.

A Implementation Details

We implement our method using greedy decoding with a maximum generation length of 128 tokens. For implementation efficiency, rather than modifying the raw image inputs corresponding to $\mathbf{v}_{\text{masked}}$, we mask the attention weights associated with visual tokens in $\mathbf{v}_{\text{top}}$ when computing IS_{core}, following the attention knockout strategy (Geva et al., 2023; Kaduri et al., 2025). The weighting parameter α and the proportion of masked image patches K used to compute the VAUQ score are selected based on a held-out validation set, as described in Appendix F. The layer index range (l_s, l_e) is chosen heuristically based on empirical observations, as illustrated in Figure 3. For all experiments, we report results averaged over three random seeds. All experiments are conducted using Python 3.11.11 and PyTorch 2.6.0 (Paszke et al., 2019) on a single NVIDIA A100 GPU with 80GB of memory.

Ground-truth labeling. For free-form visual question answering datasets—MMVet, Visual-CoT, and ViLP—we employ GPT-5 (OpenAI, 2025) as an evaluator under the *LLM-as-a-judge* paradigm (Zheng et al., 2023) to annotate model outputs. Specifically, we assess the correctness of LVLM-generated responses by determining their semantic equivalence to the corresponding gold-standard answers. We use the following evaluation prompt:

Input prompt for GPT-5-based labeling

Prompt:
Ground truth: {ground_truth}.
Model answer: {model_answer}.
Please verify whether the model answer matches the ground truth. Respond with either `Correct` or `Wrong` only.

A response is labeled as correct if the judge outputs `Correct`, and as a hallucination otherwise. To improve label consistency, we sample three independent judgments and assign the final label via majority voting. For the multiple-choice dataset CVBench, we use exact answer matching for labeling.

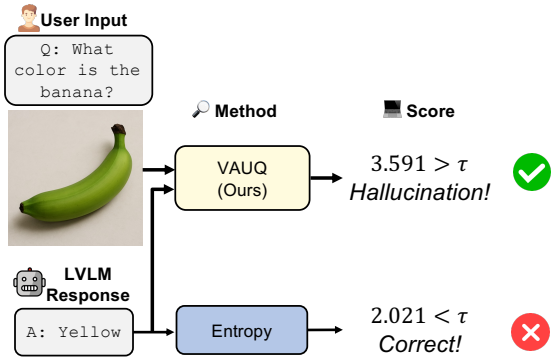


Figure 6: Qualitative example comparing our method with the entropy baseline.

B Qualitative Analysis

B.1 Qualitative Example

In Figure 6, we show a qualitative example comparing our method with the entropy baseline. The entropy baseline produces a lower uncertainty score than the threshold (τ), predicting the generated response as correct, which illustrates the language prior problem inherent in LLM-based self-evaluation. In contrast, VAUQ explicitly accounts for visual information and produces a higher uncertainty score, accurately identifying the hallucinated output.

B.2 Case Studies of Core Region Masking

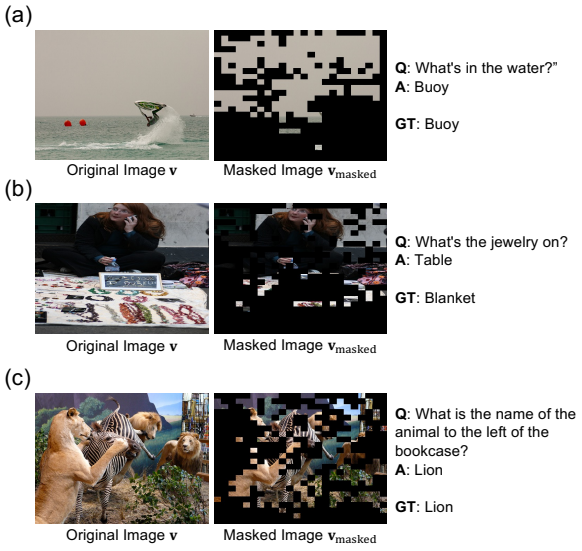


Figure 7: Qualitative case studies of core-region masking using LLaVA-1.5-7B, illustrating (a) correct response, (b) incorrect response, and (c) failure mode.

We conduct a qualitative analysis of core region masking under three scenarios: (a) correct

response, (b) incorrect response, and (c) failure mode analysis, as shown in Figure 7. We visualize the masked images using visual attention from the 10th–25th layers of LLaVA-1.5-7B, with $K = 60$. For (a), when the response is correct, the model successfully identifies and masks out semantically meaningful regions. For (b), although the model correctly captures meaningful regions of the image, it fails to reason about the output, indicating that accurate region identification alone is insufficient for correct reasoning. In this case, VAUQ captures information more effectively, as reflected by its performance compared to SVAR and our method in Tables 1 and 2. For (c), we observe a failure case in which multiple dominant objects are present in the image, causing visual attention to miss some relevant information.

C Dataset Details

ViLP (Luo et al., 2025) is a free-form vision–language evaluation dataset designed to probe language priors in LVLMs. It contains 300 carefully constructed questions, each paired with three distinct image–answer sets: one factual answer that can be inferred from language context alone, and two counterfactual answers that require grounded visual reasoning to answer correctly, yielding 900 Question–Image–Answer (QIA) triplets in total. In our work, we use the paired factual–counterfactual subset comprising 600 QIA triplets.

MMVet (Yu et al., 2024) is a vision–language evaluation benchmark designed to assess the integrated vision–language capabilities of LVLMs. It defines six core vision–language capabilities—recognition, OCR, knowledge, spatial awareness, language generation, and math—and evaluates 16 capability integrations derived from their combinations. MM-Vet consists of 218 open-ended questions paired with images, where each question requires one or more capabilities to answer correctly.

VisualCoT (Shao et al., 2024) is a free-form vision–language dataset designed to support and evaluate visual chain-of-thought reasoning in multimodal large language models. Each example is annotated with a question, answer, and an intermediate bounding box that highlights the key evidence image region required for reasoning. The dataset spans multiple domains, including text/document understanding, charts, general VQA, fine-grained recognition, and relational reasoning. In our experiments, we sample 1.5K examples from the full VisualCoT dataset for evaluation.

CVBench (Tong et al., 2024) is a multiple-choice, vision-centric evaluation benchmark designed to assess fundamental visual understanding in LVLMs. It repurposes classic vision benchmarks into vision–language questions that probe core 2D and 3D perception abilities, including spatial relationships, object counting, depth order, and relative distance. CVBench consists of 2,638 manually inspected examples.

D Additional Analysis

D.1 Evaluation on HallusionBench

To further validate our method, we evaluate on HallusionBench (Guan et al., 2024) using LLaVA-1.5-7B and Qwen-2.5-VL-7B with the same baselines as in the main experiments. As shown in Table 6, our method outperforms multi-sampling baselines, surpassing VL-Uncertainty by +1.9% AUROC on LLaVA-1.5-7B and Semantic Entropy by +0.3% on Qwen-2.5-VL-7B, and consistently outperforms all baselines across both models.

Method	LLaVA	Qwen
Perplexity	56.4	69.4
Verbalized	49.1	57.9
SVAR	60.2	63.3
Contextual Lens	54.9	58.3
Chain-of-Embeddings	52.1	61.0
Semantic Entropy	64.3	74.0
VL-Uncertainty	65.1	73.5
Ours	67.0	74.3

Table 6: Experiment results on HallusionBench.

D.2 Comparison with Other Masking Strategies

We compare our attention-based core-region masking against two alternative strategies: an embedding-based baseline (Contextual Lens) and a Grad-CAM–based (Selvaraju et al., 2020) masking variant.

Localization Quality. To quantify localization quality, we measure the overlap between attention-derived core regions and ground-truth object masks on ImageNet-S (Gao et al., 2022), which provides pixel-level annotations. As shown in Table 7, our attention-based masking achieves consistently

stronger alignment with ground-truth object regions than the embedding baseline across all metrics.

Method	Pixel Acc.↑	mIoU↑	mAP↑
Embedding (Contextual Lens)	50.4	36.1	53.9
Attention (Ours)	**69.3**	**46.4**	**77.1**

Table 7: Localization quality on ImageNet-S.

We further report category-wise IoU across 10 sampled object categories spanning varied object sizes in Table 8.

Category	Embedding	Attention (Ours)
Monkey	47.62	50.79
Hamster	31.36	38.52
Car	36.35	43.46
Iron	41.47	41.89
Pasta	67.52	70.10
Sign	15.35	20.95
Sandals	29.19	35.53
Pen	12.93	14.87
Ship	20.96	24.70
Water tower	16.60	23.84

Table 8: Category-wise IoU on ImageNet-S.

Uncertainty Quantification Performance. Since applying CAM-style methods (Zhou et al., 2016) directly to autoregressive LVLMs is non-trivial due to their multi-token generative outputs, we implement a Grad-CAM–based (Selvaraju et al., 2020) masking variant by aggregating gradients over the generated answer tokens to obtain a saliency map over image patches. Table 9 compares this gradient-based masking with our attention-derived masking on LLaVA-1.5-7B.

Method	ViLP	VisualCoT
Grad-CAM	76.0	76.6
Attention (Ours)	**77.0**	**77.8**

Table 9: AUROC comparison of masking strategies on ViLP and VisualCoT.

D.3 Evaluation with AUPRC

AUROC alone may not fully capture performance under class-imbalanced conditions. To address this, we additionally report AUPRC, which is more sensitive to class distribution. Table 10 presents results on ViLP with LLaVA-1.5-7B, where the ratio of correct to incorrect samples is approximately 2:3. Our method achieves a +8.0% AUPRC improvement over Semantic Entropy and outperforms all baselines on both metrics, demonstrating robustness under class-imbalanced settings.

Method	AUROC (%)↑	AUPRC (%)↑
Perplexity	54.6	50.5
Verbalized	56.3	47.8
SVAR	50.6	47.4
Contextual Lens	56.7	52.1
Chain-of-Embeddings	52.0	45.9
EigenScore	63.2	59.7
Semantic Entropy	63.7	60.2
VL-Uncertainty	55.6	54.9
Ours	**77.0**	**68.2**

Table 10: AUROC and AUPRC comparison on ViLP with LLaVA-1.5-7B.

E Related Works

E.1 Baselines

Perplexity (Ren et al., 2022) measures the uncertainty of a model over a generated sequence and is defined as the exponentiated average negative log-likelihood:

$$s_{\text{ppl}} = \exp\left(-\frac{1}{M}\sum_{j=1}^{M}\log p(y_j \mid \mathbf{y}_{<j}, \mathbf{v}, \mathbf{t})\right). \quad (7)$$

Verbalized Confidence (Kadavath et al., 2022) estimates model uncertainty by prompting a large language model to explicitly report its confidence in a given answer:

Input prompt for verbalized confidence

Prompt:
Question: {question}.
Model answer: {model_answer}.
On a scale of 0 to 100, how confident are you about the correctness of this answer? Respond with only a single number.

Summed Visual Attention Ratio (Jiang et al., 2025). The Visual Attention Ratio (VAR) quantifies the extent to which a generated token y_j attends to visual information by summing the attention weights assigned to image tokens within a specific attention head h and layer ℓ:

$$\text{VAR}^{(\ell,h)}(y_j) \triangleq \sum_{i=1}^{N} A^{(\ell,h)}(y_j, v_i), \quad (8)$$

where $A^{(\ell,h)}(y_j, v_i)$ denotes the attention weight from the generated token y_j to the i-th image token v_i at the h-th attention head of the ℓ-th layer, and N is the number of image tokens.

Building on this definition, the Summed Visual Attention Ratio (SVAR) aggregates visual attention by averaging VAR across all attention heads and

summing over a selected range of layers. Specifically, for token y_j, SVAR is computed over layers $\ell = 5$ to 18 as

$$s_{\text{SVAR}}(y_j) = \frac{1}{H} \sum_{\ell=5}^{18} \sum_{h=1}^{H} \text{VAR}^{(\ell,h)}(y_j), \quad (9)$$

where H denotes the total number of attention heads. The final SVAR score s_{SVAR} is obtained by averaging $s_{\text{SVAR}}(y_j)$ over all generated tokens. To align this metric with other uncertainty scores, we negate the score when reporting results.

Contextual Lens (Phukan et al., 2025) measures the alignment between textual and visual representations by computing the maximum cosine similarity between the averaged hidden representation of the generated description at layer l_T and each image token representation at layer l_I:

$$s_{\text{CL}} = \max_{i \in [N]} \text{sim}\left(\frac{1}{M} \sum_{j=1}^{M} h_{l_T}(y_j),\, h_{l_I}(v_i) \right). \quad (10)$$

To align this metric with other uncertainty scores, we negate the similarity value when reporting results.

Chain-of-Embeddings (Wang et al., 2025b) estimates response correctness by analyzing the latent trajectory of hidden states produced during inference. Specifically, it treats the sequence-level hidden representations across layers as a latent thinking path and measures its geometric variation, which differs systematically between correct and incorrect responses. A representative CoE score is defined by aggregating layer-wise changes between adjacent hidden states:

$$s_{\text{CoE}} = \frac{1}{L} \sum_{\ell=0}^{L-1} \left(\|h_{\ell+1} - h_\ell\|_2 \; - \; \arccos \frac{h_{\ell+1}^\top h_\ell}{\|h_{\ell+1}\| \|h_\ell\|} \right), \quad (11)$$

where h_ℓ denotes the hidden embedding at layer ℓ, averaged over all generated tokens.

EigenScore (Chen et al., 2024) measures uncertainty by quantifying the divergence among multiple generated responses in the model's internal embedding space. Given K sampled responses, hidden embeddings are extracted from internal states and used to form a covariance matrix, whose eigenvalues capture semantic diversity. The EigenScore is defined as the average log-determinant of the covariance matrix:

$$s_{\text{Eigen}} = \frac{1}{K} \sum_{i=1}^{K} \log(\lambda_i), \quad (12)$$

where $\{\lambda_i\}_{i=1}^{K}$ are the eigenvalues of the regularized covariance matrix of sentence embeddings. Higher scores indicate greater semantic divergence and higher uncertainty.

Semantic Entropy (Kuhn et al., 2023) measures uncertainty over meanings rather than surface forms by accounting for semantic equivalence among generated responses. Given a set of sampled generations clustered into semantic equivalence classes $\mathcal{C}$, semantic entropy is defined as the entropy of the induced distribution over meanings:

$$s_{\text{SE}} = -\sum_{c \in \mathcal{C}} p(c \mid x) \, \log p(c \mid x), \quad (13)$$

where $p(c \mid x) = \sum_{s \in c} p(s \mid x)$ aggregates the probabilities of all sequences s that share the same semantic meaning.

VL-Uncertainty (Zhang et al., 2025) estimates uncertainty in LVLMs by measuring the variability of model responses to semantically equivalent perturbations of both visual and textual prompts. Specifically, multiple perturbed image–text pairs are constructed, the corresponding responses are clustered by semantic meaning, and uncertainty is quantified as the entropy of the resulting cluster distribution:

$$s_{\text{VL-U}} = -\sum_{c \in \mathcal{C}} p(c) \log p(c), \quad (14)$$

where $\mathcal{C}$ denotes the set of semantic answer clusters and $p(c)$ is the proportion of responses in cluster c. Higher entropy indicates greater uncertainty and a higher likelihood of hallucination.

E.2 Comparison with Prior Work

Novelty and positioning relative to VCD. While both VCD and our method involve contrasting distributions, their goals and mechanisms differ fundamentally. VCD is a decoding-time hallucination mitigation method that modifies token generation by contrasting logits from original vs. distorted images to reduce hallucinations during generation. In contrast, VAUQ is a post-hoc self-evaluation and hallucination detection framework that produces an uncertainty score estimating answer correctness by measuring how predictive uncertainty (e.g., entropy) changes when core visual evidence is masked.

Beyond simple contrastive comparison, we introduce a core-region masking strategy derived

from intermediate attention maps, requiring no additional training or labels. We show that masking only the core visual regions is more effective than masking the entire image, as in VCD, which applies global visual noise, and that combining mid-layer visual signals with output-level entropy provides a robust uncertainty signal for LVLM self-evaluation.

Distinction from attention-guided erasing methods. Our method is training-free and operates on top of a pre-trained autoregressive LVLM by leveraging its own self-attention signals to identify core regions. In contrast, prior approaches (Liu et al., 2019; Shi et al., 2022) learn a masking module or introduce additional components during training. We further show that intermediate layers yield more reliable signals for unsupervised localization than final-layer features, which directly informs our masking design.

Unlike methods that aim to improve grounding performance (Liu et al., 2019) or learned representations (Shi et al., 2022), VAUQ is primarily a self-evaluation framework. Core-region masking serves only as a mechanism to measure how model confidence changes when key visual evidence is removed, rather than to modify training dynamics.

Finally, combining intermediate-layer attention-based masking with output-level predictive entropy yields a unified uncertainty score tailored for response-level hallucination detection in LVLM self-evaluation. The contribution thus lies not only in the masking mechanism itself, but in its integration into a principled uncertainty quantification framework for reliability assessment.

E.3 Extended Literature Review

Large Vision-Language Models (LVLMs) have demonstrated remarkable capabilities in understanding the visual world by modeling interactions between visual and textual modalities (Dai et al., 2023; Tong et al., 2024; Wang et al., 2025a; Hong et al., 2025; Kimi Team, 2025). These models integrate a vision encoder (Radford et al., 2021) with a language model (Grattafiori et al., 2024) via various multimodal fusion modules (e.g., MLPs). Through visual instruction tuning, LVLMs perform complex image understanding and reasoning tasks, laying the foundation for a wide range of applications, including agentic (Durante et al., 2024) and embodied systems (Kim et al., 2024). Despite these advances, LVLMs remain prone to hallucinations, posing significant risks for real-world deployment.

Object Hallucinations in LVLMs refer to cases where the model generates plausible-sounding mentions of objects that are not present in the image. In contrast to higher-level response hallucinations, detecting and mitigating such object-level errors has been an active area of research (Liu et al., 2024b). Existing approaches can be broadly categorized into training-based and training-free approaches. Training-based methods typically involve additional supervision or model updates to strengthen vision–language alignment, such as fine-tuning with hallucination-aware objectives (Sun et al., 2024; Lu et al., 2025). On the other hand, training-free methods seek to mitigate hallucinations at inference time, for example, by modifying decoding strategies using contrastive decoding (Leng et al., 2024; Favero et al., 2024; Liu et al., 2024c) or latent space steering (Liu et al., 2025; Duan et al., 2025; Yang et al., 2025; An et al., 2025).

In contrast to these object-level hallucinations, our method focuses on response-level self-evaluation (i.e., hallucination detection), aiming to capture broader forms of implausible generation beyond individual object errors.

F Hyperparameter settings

Dataset	Hyperparameters		
	(l_s, l_e)	α	K
ViLP	(10,25)	0.6	60
MMVet		0.6	40
VisualCoT		0.3	60
CVBench		1.2	30

Table 11: Hyperparameter setting for LLaVA-1.5-7B.

Dataset	Hyperparameters		
	(l_s, l_e)	α	K
ViLP	(10,35)	1.5	20
MMVet		0.4	30
VisualCoT		0.2	60
CVBench		1.2	40

Table 12: Hyperparameter setting for LLaVA-1.5-13B.

G Additional Future Works

Our experiments focus on a subset of widely used, instruction-tuned LVLMs and on image-based benchmarks. This focus is important because

Dataset	Hyperparameters		
	(l_s, l_e)	α	K
ViLP	(12,26)	0.8	80
MMVet		0.1	60
VisualCoT		0.1	50
CVBench		2.0	60

Table 13: Hyperparameter setting for Qwen2.5-VL-7B.

Dataset	Hyperparameters		
	(l_s, l_e)	α	K
ViLP	(10,25)	0.5	70
MMVet		0.1	50
VisualCoT		0.1	50
CVBench		0.2	60

Table 14: Hyperparameter setting for InternVL3.5-8B.

instruction-tuned LVLMs and image-based benchmarks constitute the dominant setting in which current multimodal systems are deployed and evaluated, and thus provide a realistic and high-impact testbed for studying practical self-evaluation under visual grounding constraints. At the same time, we acknowledge the recent and rapidly emerging trend toward more advanced reasoning-oriented LVLMs, including models designed for long chain-of-thought reasoning, complex multi-step visual inference, and video understanding. While we do not explicitly evaluate VAUQ in these settings, the core principles of vision-aware uncertainty quantification remain applicable, though extending the framework may require modeling how visual information contributes across multiple reasoning steps rather than only at the final response level. We hope our framework can serve as a strong baseline and foundation for future research in these directions.