Image and Vision Computing

Sarah: Hallucination Detection for Large Vision Language Models with Semantic Information Locator and Purifier in Uncertainty Quantification Method

--Manuscript Draft--

Manuscript Number:	
Article Type:	Full Length Article
Keywords:	Hallucination Detection; Large Vision Language Model; Semantic Information; Uncertainty Quantification; Single-round Inference
Corresponding Author:	Yang Zhang, Ph.D. Beijing Jiaotong University CHINA
First Author:	Yue Fang, Bachelor
Order of Authors:	Yue Fang, Bachelor
	Yang Zhang, Ph.D.
	Yawen Liu, Master
	Yetian Yu, Master
Abstract:	Large Vision-Language Models (LVLMs) have demonstrated remarkable potential in multi-modal applications, yet their reliability is compromised by hallucination --misalignment between generated text and visual inputs, linguistic context, or factual knowledge. To address this urgent challenge, we propose Sarah (Hallucination Detection for Large Vision Language Models with Semantic Information Locator and Purifier in Uncertainty Quantification Method), a novel hallucination detection framework grounded in uncertainty quantification method. Different from most existing uncertainty-based methods that utilize the variance of multi-round inference and need complex external tools, Sarah requires only single-round of inference result and minimal dependence on external tools, delving deeply into the value of the probability distribution. Considering uneven distribution of semantic information in both complete generation as well as possible outputs per-step, Semantic Information Locator and Purifier are proposed to enhance semantic collaboration and reduce semantic interference. Our extensive experiments across 5 off-the-shelf LVLMs and 2 open-ended visual question-answering benchmarks demonstrate that Sarah demonstrates superior performance by outperforming five out of six selected strong baseline methods in hallucination detection, while achieving comparable detection accuracy to the remaining one with significantly enhanced cost-effectiveness (requires only 1/25 of the computational time per iteration). Specifically, on image captioning outputs generated by GPT-4o, Sarah achieves a hallucination detection accuracy of 86.6%. Analysis over LVLMs further exposes critical limitations: over 13.4% of outputs from state-of-the-art LVLMs exhibit hallucination, which leaves room for future improvements.

Abstract

Large Vision-Language Models (LVLMs) have demonstrated remarkable potential in multi-modal applications, yet their reliability is compromised by hallucination -- misalignment between generated text and visual inputs, linguistic context, or factual knowledge. To address this urgent challenge, we propose Sarah (Hallucination Detection for Large Vision Language Models with Semantic Information Locator and Purifier in Uncertainty Quantification Method), a novel hallucination detection framework grounded in uncertainty quantification method. Different from most existing uncertainty-based methods that utilize the variance of multi-round inference and need complex external tools, Sarah requires only single-round of inference result and minimal dependence on external tools, delving deeply into the value of the probability distribution. Considering uneven distribution of semantic information in both complete generation as well as possible outputs per-step, Semantic Information Locator and Purifier are proposed to enhance semantic collaboration and reduce semantic interference. Our extensive experiments across 5 off-the-shelf LVLMs and 2 open-ended visual question-answering benchmarks demonstrate that Sarah demonstrates superior performance by outperforming five out of six selected strong baseline methods in hallucination detection, while achieving comparable detection accuracy to the remaining one with significantly enhanced cost-effectiveness (requires only 1/25 of the computational time per iteration). Specifically, on image captioning outputs generated by GPT-4o, Sarah achieves a hallucination detection accuracy of 86.6%. Analysis over LVLMs further exposes critical limitations: over 13.4% of outputs from state-of-the-art LVLMs exhibit hallucination, which leaves room for future improvements.

Highlights

Sarah: Hallucination Detection for Large Vision Language Models with Semantic Information Locator and Purifier in Uncertainty Quantification Method

Fang Yue, Zhang Yang, Liu Yawen, Yu Yetian

- We propose a hallucination detection method grounded in uncertainty quantification, which achieves performance comparable to state-of-the-art approaches while being significantly more cost-effective, as it avoids multiple inference rounds and complex external tools.
- Two novel modules take into account the uneven distribution of semantic information, realizing semantic cooperation strengthen and semantic interference mitigation.
- The versatile framework is applicable to various large vision language models and generation tasks.

Sarah: Hallucination Detection for Large Vision Language Models with Semantic Information Locator and Purifier in Uncertainty Quantification Method

Fang Yue[a], Zhang Yang[b,*], Liu Yawen[c] and Yu Yetian[d]

[a] School of Electronic and Information Engineering, Beijing Jiaotong University, Beijing, 100044, China
[b] School of Cyber Science and Engineering, Southeast University, Nanjing, 214135, China

ARTICLE INFO

Keywords:
Hallucination Detection
Large Vision Language Model
Semantic Information
Uncertainty Quantification
Single-round Inference

ABSTRACT

Large Vision-Language Models (LVLMs) have demonstrated remarkable potential in multi-modal applications, yet their reliability is compromised by hallucination – misalignment between generated text and visual inputs, linguistic context, or factual knowledge. To address this urgent challenge, we propose Sarah (Hallucination Detection for Large Vision Language Models with Semantic Information Locator and Purifier in Uncertainty Quantification Method), a novel hallucination detection framework grounded in uncertainty quantification method. Different from most existing uncertainty-based methods that utilize the variance of multi-round inference and need complex external tools, Sarah requires only single-round of inference result and minimal dependence on external tools, delving deeply into the value of the probability distribution. Considering uneven distribution of semantic information in both complete generation as well as possible outputs per-step, Semantic Information Locator and Purifier are proposed to enhance semantic collaboration and reduce semantic interference. Our extensive experiments across 5 off-the-shelf LVLMs and 2 open-ended visual question-answering benchmarks demonstrate that Sarah demonstrates superior performance by outperforming five out of six selected strong baseline methods in hallucination detection, while achieving comparable detection accuracy to the remaining one with significantly enhanced cost-effectiveness (requires only 1/25 of the computational time per iteration). Specifically, on image captioning outputs generated by GPT-4o, Sarah achieves a hallucination detection accuracy of 86.6%. Analysis over LVLMs further exposes critical limitations: over 13.4% of outputs from state-of-the-art LVLMs exhibit hallucination, which leaves room for future improvements.

1. Introduction

Large Vision-Language Models (LVLMs) have represent a significant advancement in multi-modal AI systems, demonstrating the ability to process and integrate visual and textual data for solving complex tasks that require joint understanding of both modalities (Liang, Xu, Hong, Shang, Wang, Fu and Liu (2024); Liu, Xue, Chen, Chen, Zhao, Wang, Hou, Li and Peng (2024); Zhang, Huang, Jin and Lu (2024b); Jin, Li, Liu, Gu, Wu, Jiang, He, Zhao, Tan, Gan et al. (2024)). Despite their impressive capabilities, these models are prone to hallucination —— a critical issue where generated content deviates from factual accuracy or contradicts the provided visual and textual context (Liu et al. (2024); Huang, Yu, Ma, Zhong, Feng, Wang, Chen, Peng, Feng, Qin et al. (2025)). This phenomenon primarily stems from limitations such as constrained knowledge boundaries, failures in knowledge recall, and reliance on outdated pre-training corpora (Liu et al. (2024); Schulman, Wolski, Dhariwal, Radford and Klimov (2017)). Addressing hallucination detection, which involves verifying the factual consistency and accuracy of each information unit in generated outputs relative to the input image and text, presents three major challenges. First, LVLM outputs typically interweave correct and hallucinated information (Min, Krishna, Lyu, Lewis, Yih, Koh, Iyyer, Zettlemoyer and Hajishirzi (2023)), transforming the detection task into a complex, fine-grained analysis rather than a simple binary classification. Second, exhaustive validation against external references proves computationally intensive and resource-prohibitive. Third, existing evaluation metrics often produce opaque, single-dimensional scores, offering limited interpretability regarding the specific nature and location of errors.

To address these challenges, we present Sarah, a novel uncertainty-based framework for LVLM hallucination detection. Our approach introduces four key innovations: (1)Independent Claim Extraction: The framework decomposes the generated content into discrete claims, where each claim constitutes a concise statement representing an individual unit of information. (Fadeeva, Rubashevskii, Shelmanov, Petrakov, Li, Mubarak, Tsymbalov, Kuzmin, Panchenko, Baldwin et al. (2024); Hu, Liu, Kasai, Wang, Ostendorf, Krishna and Smith (2023b); Jing, Li, Chen and Du (2023)); (2)Semantic Information Processing: Through our Semantic Information Locator and Purifier modules, we enhance semantic coherence while mitigating interference. The Locator identifies semantically significant words within the global context of the output, while the Purifier processes token-level distributions, aggregates semantically equivalent tokens, and eliminates uncertainties arising from linguistic variations; (3)Uncertainty Quantification: Each extracted claim undergoes rigorous validation through our uncertainty quantification method; (4)Decision Mechanism:

[*] Corresponding author: zhang.yang@bjtu.edu.cn
ORCID(s):

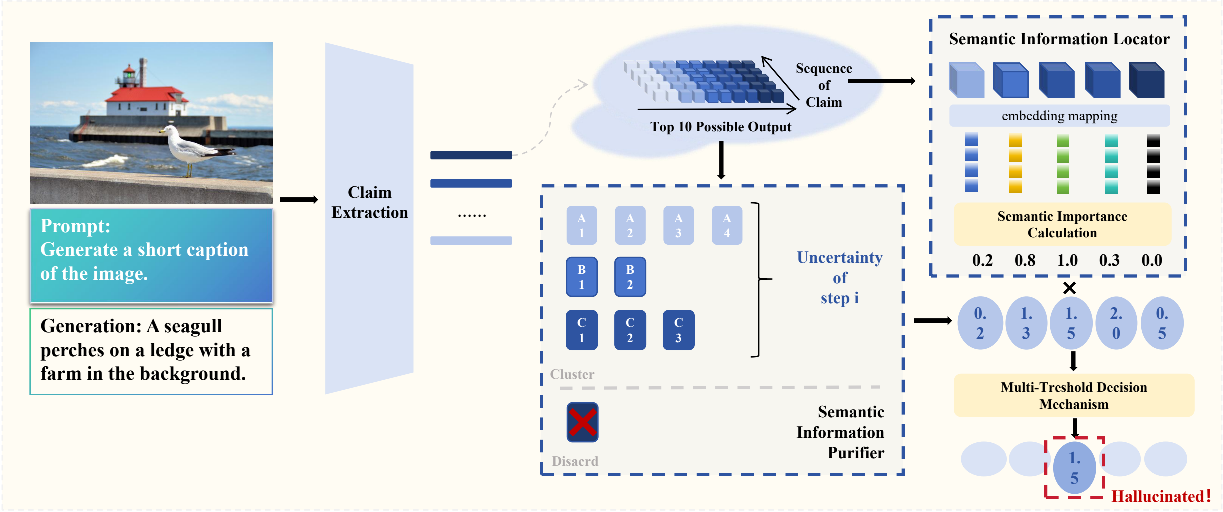

Prompt:
Generate a short caption of the image.
Generation: A seagull perches on a ledge with a farm in the background.
Claim Extraction
......
Sequence of Claim
Top 10 Possible Output
Semantic Information Locator
embedding mapping
Semantic Importance Calculation
0.2
0.8
1.0
0.3
0.0
×
A 1
A 2
A 3
A 4
B 1
B 2
C 1
C 2
C 3
Uncertainty of step i
Cluster
Disacrd
Semantic Information Purifier
0. 2
1. 3
1. 5
2. 0
0. 5
Multi-Treshold Decision Mechanism
1. 5
Hallucinated!

Figure 1: Overview of the Sarah framework for hallucination detection in LVLMs. The framework consists of four key stages: (1)Independent Claim Extraction, where atomic claims are decomposed from LVLM-generated responses for granular analysis; (2)Semantic Information Locator, which computes attention-based weights for each token to guide probability recalibration; (3)Semantic Information Purifier, where token-level probabilities are refined using part-of-speech analysis to enhance detection robustness; and (4)Multi-Threshold Decision Mechanism, which combines probabilistic and semantic criteria to classify hallucinated content. Sarah provides a detailed, interpretable, and scalable approach for evaluating hallucination severity in free-form LVLM outputs.

Our Multi-Thresholds Decision Mechanism focuses exclusively on high-uncertainty tokens, effectively minimizing interference from low-hallucination content. As illustrated in Figure.1, Sarah successfully identifies and isolates hallucinated claims, such as the second claim in the example.

The proposed detection framework offers two significant advantages: it precisely identifies the specific components that contribute to hallucination, and provides a mathematical approach for uncertainty quantification that requires neither external knowledge nor model modifications. Comprehensive evaluations across two established vision-language benchmarks demonstrate that Sarah achieves strong alignment with both human assessments and our specialized automated labeling system. (Cui, Zhou, Yang, Wu, Zhang, Zou and Yao (2023); Lin, Maire, Belongie, Hays, Perona, Ramanan, Dollár and Zitnick (2014)). Our results show substantial improvements in hallucination detection for LVLM-generated content. Furthermore, our comprehensive analysis using Sarah reveals critical insights into the reliability of various LVLMs, with GPT-4o emerging as the top performer among tested models, particularly in image captioning tasks. These findings not only corroborate human evaluations but also align with previous research, establishing Sarah as a robust tool for model assessment and improvement. In summary, the primary contributions of this work are as follows:

1.We propose Sarah, a novel cost-effective hallucination detection framework with uncertainty quantification method for LVLMs. Sarah requires only single round inference and has minimal dependence on external tools, significantly reducing computational overhead.

2.We introduce a semantic information processing module coupled with a rational decision-making mechanism within Sarah. Extensive experiments demonstrate that this integration enables Sarah to outperform over 5 out of 6 selected strong baseline methods. Moreover, while achieving comparable performance to the remaining one (InterrogateLLM), it only requires 1/25 of the time.

3.Through comprehensive empirical analysis using Sarah, we indicate that over 13. 4% of model responses contain hallucination, and identifies GPT-4o as exhibiting the lowest hallucination rates among contemporary models, particularly in image captioning tasks and mitigating OCR-related biases, providing valuable benchmarks for future model development.

2. Related Work

2.1. Large Vision Language Models

Large Vision-Language Models (LVLMs) have emerged as a transformative paradigm in multimodal AI, integrating three core architectural components: a text encoder, an image encoder, and a cross-modal alignment module (Chen, Guo, Yi, Li and Elhoseiny (2022); Huang, Dong, Wang, Hao, Singhal, Ma, Lv, Cui, Mohammed, Patra et al. (2023)). Recent advancements in LVLMs have significantly pushed the boundaries of performance across a diverse range of vision-language (VL) tasks, including autonomous driving, embodied robotics, and medical diagnosis (Moor, Huang, Wu, Yasunaga, Dalmia, Leskovec, Zakka, Reis and Rajpurkar (2023); Peng, Wang, Dong, Hao, Huang, Ma and Wei (2023); Wang, Xie, Hu, Zou, Fan, Tong, Wen, Wu, Deng, Li et al. (2023b); Cui, Ma, Cao, Ye, Zhou, Liang, Chen, Lu, Yang, Liao et al. (2024)).

Notable contributions in this domain include LLaVA (Park, O'Brien, Cai, Morris, Liang and Bernstein (2023)), which bridges visual and textual modalities by projecting visual encoder outputs into the input space of the LLaMA language model (Touvron, Lavril, Izacard, Martinet, Lachaux, Lacroix, Rozière, Goyal, Hambro, Azhar et al. (2023)) and training on synthetic multimodal data (Leng, Xing, Cheng, Zhou, Zhang, Li, Zhao, Lu, Miao and Bing (2024)). mPLUG-Owl (Strohauer, Wietschorke, Zugliani, Flaschmann, Schmid, Grotowski, Müller, Jonas, Althammer, Gross et al. (2023)) further advances multimodal capabilities through a multitask learning framework, enhancing cross-modal understanding. GPT-4o (OpenAI (2023)) achieves finer-grained image comprehension by refining its visual encoder and optimizing the alignment of visual features with its language model via an enhanced adapter. Most recently, Llama-v3.2-vision (Stock, Schlögl and Groth (2023)) has set new benchmarks in generalization and complex VL task performance by incorporating more efficient visual attention mechanisms and leveraging a larger-scale, diverse dataset.

Despite their remarkable capabilities, LVLMs inherit the hallucination issue prevalent in Large Language Models (LLMs), posing significant challenges to AI safety and reliability. To address this critical limitation, we propose an intrinsic hallucination detection method designed to enhance the safety and trustworthiness of human-LVLM interactions.

2.2. Hallucination on LVLMs

Hallucination in LVLMs refers to the generation of content that is factually inconsistent with the provided visual and textual context. This phenomenon arises from various factors, including the presence of noisy or incomplete training data (Liu et al. (2024); Hu, Zhang, Zhao and Sun (2023a); Liu, Lin, Li, Wang, Yacoob and Wang (2023a)), suboptimal attention allocation during training (Mao, Lu, Zhang and Wang (2023)), imperfect reasoning strategies, or inadequate decoding prompts (Liu et al. (2024)). While several methods have been proposed to detect and evaluate hallucinations, existing approaches are often limited to specific types of hallucination or constrained answer formats, failing to capture the full spectrum of hallucination phenomena. Moreover, current hallucination detection techniques frequently achieve high accuracy at the cost of significant resource consumption: they heavily rely on external tools and often require multi-round inference.

For instance, CHAIR (Rohrbach, Hendricks, Burns, Darrell and Saenko (2018)) quantifies object hallucination by comparing generated captions with ground-truth objects in images using binary "yes-or-no" questions. Building on this, Li et al. (Li, Du, Zhou, Wang, Zhao and Wen (2023)) introduced POPE, a polling-based query technique for object hallucination detection. More recently, researches (Min et al. (2023); Jing et al. (2023); Liu, Lin, Li, Wang, Yacoob and Wang (2023b)) have expanded the scope of hallucination evaluation beyond single-instruction formats and human-annotated ground truth, leveraging external tools such as advanced models to quantify hallucinations in free-form outputs.

In contrast to these approaches, our work focuses on reference-free hallucination detection method for LVLMs, eliminating the need for external tools or multi-round inference. Our method is designed to handle open-ended visual question-answering settings, where answers are free-form and potentially lengthy, addressing a broader range of hallucination types.

2.3. Uncertainty-Based Reference-Free Methods for Hallucination Detection

Reference-free hallucination detection methods, which rely solely on internal data of the model, have gained significant traction in LLM researchLi, Geng, Lyu, Zhu, Panov and Karray (2024). Among these, uncertainty-based methods (Fadeeva et al. (2024); Kadavath, Conerly, Askell, Henighan, Drain, Perez, Schiefer, Hatfield-Dodds, DasSarma, Tran-Johnson et al. (2022); Li et al. (2024); Kuhn, Gal and Farquhar (2023); Zhang, Zhang and Zheng (2024c)) are particularly promising, operating on the hypothesis that generation with high uncertainty is more likely to be hallucinated (Wang, Zhou, Xu, Shi, Zhao, Xu, Ye, Yan, Zhang, Zhu et al. (2023a)). These methods have been extensively studied in LLMs and are increasingly being adapted to LVLMs.

Existing uncertainty-based approaches can be broadly categorized into three paradigms: (1) consistency-based methods, which sample multiple responses to the same input and evaluate the consistency of factual statements (Duan, Cheng, Wang, Zavalny, Wang, Xu, Kailkhura and Xu (2023); Manakul, Liusie and Gales (2023)); (2)input perturbation methods, which assess prediction variance by perturbing the original input (Zhang et al. (2024c); Zhang, Gao, Jiang, Zhao and Zheng (2024a)); and (3)Bayesian methods, which quantify uncertainty by feeding the same input multiple times into stochastic networks with dynamic weights (Blundell, Cornebise, Kavukcuoglu and Wierstra (2015); Gal and Ghahramani (2016); Kendall and Gal (2017)).

While these methods often require multiple rounds of model inference, we argue that a single round of output contains sufficient information for uncertainty quantification. Our work focuses on analyzing the probability distribution of tokens within a single-round output, significantly reducing computational overhead and API call requirements while maintaining robust hallucination detection capabilities.

3. Method

Approaches to hallucination evaluation and detection might rely too much on external tools, treating the LVLM as a black-box or alternatively focus on the probability distribution without accounting for the special characteristics of language (e.g., the uneven distribution of semantic information).

Our reference-free approach instead uses the powerful tools of probabilistic modeling, but also recognizes the unique challenges posed by open-ended visual question-answering tasks. In this section, we begin by clearly introducing the foundation of Sarah that quantifies the abstract concept of hallucination based on uncertainty quantification in Section 3.1, followed by a detailed framework of Sarah with modules for semantic collaboration enhancement and semantic interference mitigation in the remaining sections.

3.1. Hallucination Evaluation Based on Uncertainty Quantification

We can formalize uncertainty of generation mathematically. Let the probability distribution of the token z_i in the i-th generation step be $P(z_i, k \mid s < i, x)$ $(1 \leq i \leq N, 1 \leq k \leq K)$, where k is the k-th most likely generation at i-th generation step of sentence s, and x stands for the input. For the majority of sentences, the top k most likely outputs account for the majority of possibility, and is enough for indicating the uncertainty, see Figure.2 .

The total uncertainty of a prediction can then be acquired according to the predictive entropy of the output distribution. This measures the information one has about the output given the input. The predictive entropy for token z_i is the conditional entropy of the output random variable at step i:

$$U_{z_i} = -\sum_{k=1}^{K} p(z_{i,k} \mid s_{<i}, x) \log p(z_{i,k} \mid s_{<i}, x) \tag{1}$$

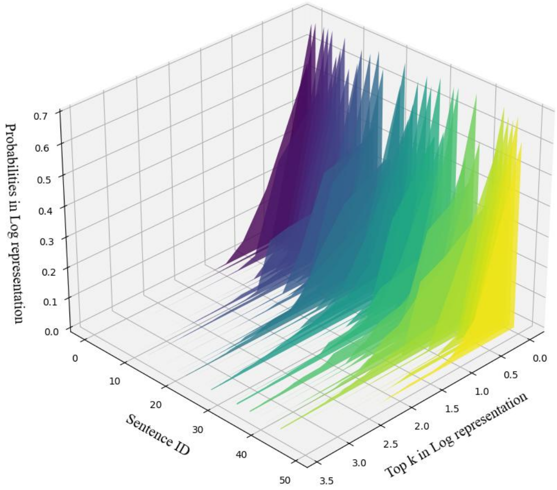


Figure 2: Probability distribution of the top k generation of each token.

3.2. Independent claim extraction

LVLM-generated text often consists of complex, lengthy sentences, making direct evaluation of hallucination prone to error omission. To address this, we decompose the generation into independent claims—basic units that convey complete meanings independently—using specially designed prompts on GPT-4. Formally, let c_s, i denote the i-th independent claim in a sentence s. This granular decomposition enables precise identification of hallucinated information.

3.3. Semantic Information Locator from Horizontal and Purifier from Vertical

Generated tokens contribute unequally to the semantic meaning of a sentence (see Figure.3), yet traditional uncertainty estimation methods treat them uniformly (Duan et al. (2023)). Besides, LVLMs often generate lexically diverse but semantically equivalent outputs, treating such variations as distinct entities might overestimate uncertainty, not to mention the interference caused by over-interpreting uncertainties in semantic style and word order. To address the former situation, we introduce the Semantic Information Locator to quantifies the semantic importance of each token and re-weights its hallucination score accordingly. To address the latter one, we introduce the Semantic Information Purifier, which groups outputs by underlying meaning, inspired by Semantic Entropy (Kuhn et al. (2023)).

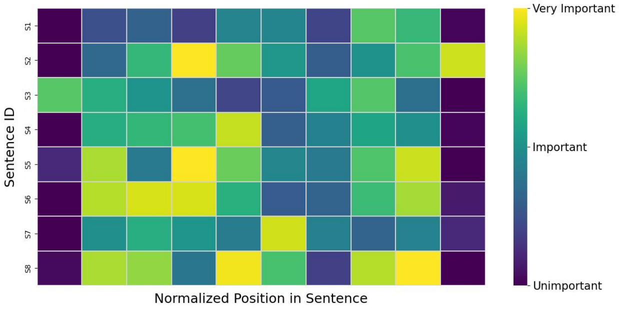


Figure 3: The uneven distribution of word semantic importance.

Semantic Information Locator. Semantic Information Locator consists of the creation of semantic fluctuation information, the mapping of semantic information for each sentence, and the calculation of sentence similarity. In practice, we first perturb the original output at step i multiple times by employing part-of-speech-equivalent perturbations without altering the underlying part of speech. To achieve this, we rephrase the original prediction utilizing words from the Universal Independent Corpus (De Marneffe, Manning, Nivre and Zeman (2021)), while preserving its narrative and structure. During each perturbation, analogous to the perturbations:

$$T_{z_{i,j}} = \varnothing(T, \tau_j), \tag{2}$$

where $j = 1, 2, \ldots, N$, and N is the number of perturbation times. T represents the initial prediction, τ_j denotes the randomness of sampling during the j-th perturbation. $\varnothing$ refers to the utilized random sampling and $T_{z_{i,j}}$ is the j-th perturbed prediction at token z_i. We map sentences that have perturbations at position i into semantic space following the BertScore approach (Moor et al. (2023)) and aggregate them into a unified set:

$$\{\langle \Gamma(T_{z_{i,0}}), \Gamma(T_{z_{i,1}}), \Gamma(T_{z_{i,j}}), ..., \Gamma(T_{z_{i,N}}) \rangle | j = 1, 2, ..., N\}, \tag{3}$$

where Γ() denotes the operation of mapping that shapes each sentence into contextual embeddings. The Semantic Importance (SI) of token z_i in the independent claim c is defined as:

$$SI(c, z_i) = 1 - \frac{1}{M} \sum_{j=1}^{M} \cos\langle \Gamma(T_{z_{i,j}}),\ \Gamma(T_{z_{i,0}}) \rangle, \qquad (4)$$

where cos<.,.> denotes the cosine similarity and j= 1, 2, . . . , N. SI range between [0, 1], with higher values indicating greater semantic impact. Tokens with negligible SI are ignored, as their uncertainty is unlikely to contribute to hallucinations.

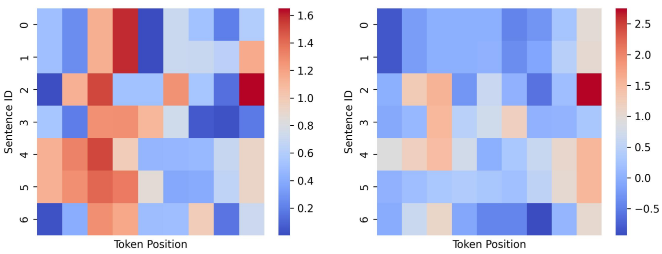


Figure 4: The distribution of sentence hallucination scores before (left) and after (right) reweighting. It can be observed that after reweighting, the values at some positions with higher uncertainty are flattened due to their low semantic importance to the sentence. This demonstrates that the semantic locator primarily functions to suppress parts with high uncertainty and low semantic importance.

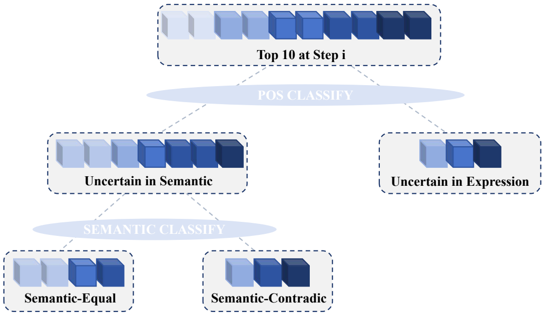


Figure 5: The decision tree of the classification procedure. Words at step i are firstly divided depending on their part of speech relationship with the exact output. Because uncertainty in expression often presents as having words organized in different expression order. The second stage of the decision tree classifies words based on their semantic relationship.

Semantic Information Purifier. Our method classify uncertainty in the generation into two types: (1) semantic uncertainty, which reflects discrepancies in conveyed information and may lead to hallucinations, and (2) expression uncertainty, which pertains to variations in language style or word order and does not make produced text less factual. To distinguish these, we leverage a pre-trained text-only LLM, prompting it to classify the top-10 words at each step based on their part-of-speech relationship and semantic relationship, according to WordNet [45]. The procedure of classification includes two steps. First, the pre-trained LLM makes a binary classification judgement on whether the uncertainty is caused by semantic meaning or expression order and style. Then, for the words tagged as semantic information uncertainty, they are further dived into two groups: semanitc-equal or semantic-contradict.

Consequently, the word uncertain in expression is discarded by setting their probabilities to zero:

$$p(z_{i,k} \mid z_{i,k} \text{ is expression uncertainty}, s_{<i}, x) = 0. \quad (5)$$

As for the words satisfying the condition of sharing the same meaning (tagged as semantic-equal), we cluster their probabilities and thus obtain a set of semantic clusters at position i, where M_e is the total number of semantic clusters. In this way various surface forms with a similar meaning are mapped to a single categorical variable. Then, recall from Eq. (1) that estimating the uncertainty by predictive entropy, the uncertainty of token z_i can be written as follows:

$$U_{\text{Purified}}(c, z_i) = -\sum_{j=1}^{M_e} p(\mathbf{e}_{i,j}) \log p(\mathbf{e}_{i,j}). \qquad (6)$$

This helps to purify the probability distribution at each generation step, making it more conducive to hallucination quantification.

3.4. LVLMs Hallucination Detection with Multi-threshold Decision Mechanism

Sarah outputs hallucination score with a continuous distribution, where the hallucination score for each token z_i in claim c is obtained by taking the product of semantic importance (SI) and purified uncertainty (U_{Purified}) proposed in Section 3.1:

$$\text{hallucination\_score}(z_i) = U_{\text{Purified}}(c, z_i) \cdot e^{(1-\alpha)} \cdot SI(c, z_i), \qquad (7)$$

here the adjustable variable serves as a trick that transforms Sarah into a flexible, adaptive evaluation method, and can be adjusted to change the level of involvement of different modules.

Inspired by the observation that low-hallucination tokens introduce noise for hallucination quantification while high-hallucination tokens dominate the overall judgment, our method propose a Multi-Threshold Decision Mechanism: Threshold 1 identifies highly hallucinatory tokens at each step:

$$\text{hallucination}(c, z_i) = \begin{cases} \text{'high'}, & \text{score}(z_i) > threshold1, \\ \text{'low'}, & \text{else.} \end{cases} \qquad (8)$$

Threshold 2 determines whether a claim contains hallucination based on the sum of scores from highly hallucinatory tokens. If the sum exceeds threshold 2, the claim is marked as hallucinated; otherwise, it is considered non-hallucinated:

$$\text{score}(c) = \begin{cases} 1 & \text{if sum(high hallucinatory)} > \text{threshold 2}, \\ 0 & \text{otherwise.} \end{cases} \tag{9}$$

The sentence-level hallucination score for the entire output is the sum of scores across all claims. Finally, the Youden Index is employed to establish an optimal threshold for detecting whether hallucination exists in the generation. The Youden Index is a statistical metric that balances sensitivity (i.e. True Positive Rate) and specificity (i.e.True Negative Rate), defined as:

$$\text{Youden Index} = \text{TPR} + \text{TNR} - 1. \tag{10}$$

The index ranges from 0 to 1, where a higher value indicates better discriminative ability. Threshold 3 is selected by maximizing Youden Index across all possible classification cutoffs, ensuring an equilibrium between correctly identifying true positives and minimizing false positives.This mechanism ensures robust and interpretable hallucination detection while minimizing the impact of low-severity errors.

4. Experiment

Baselines. Given that both LVLMs and LLMs are inherently free-form generative models, hallucination detection methods originally developed for LLMs can be effectively adapted for LVLMs with appropriate modifications. Consequently, we incorporated six up-to-date hallucination detection methods designed for either LLMs or LVLMs in our experimental framework. (1)Faithfulness to Atomic Image Facts Score(Faithscore). Faithscore extracts a comprehensive list of atomic facts from sub-sentences, and conducts consistency verification between fine-grained atomic facts and the input image. (2)GPT4-Assisted Visual Instruction Evaluation(GAVIE). GAVIE applies GPT4 to take the dense captions with bounding box coordinates as the image content and thus compare the consistency between input and output. (3)InterrogateLLM (Yehuda, Malkiel, Barkan, Weill, Ronen and Koenigstein (2024)). InterrogateLLM employs LLMs to reconstruct queries based on generated answers, subsequently identifying potential hallucinations by analyzing consistency between the original queries and the reconstructed queries derived from model outputs. (4)HalDet-LLaVA (Chen, Wang, Xue, Zhang, Yang, Li, Shen, Liang, Gu and Chen (2024)). HalDet-LLaVA is a hallucination detection model for LVLMs trained on MHaluBench. While capable of identifying hallucinations in both image-to-text and text-to-image generation tasks, our experiment focuses exclusively on the former scenario. (5)VL-Uncertainty. VL-Uncertainty identifies hallucinations in generations by quantifying the model's uncertainty regarding perturbed input data. (6)Semantic Entropy (SE). SE is firstly proposed for LLMs, which introduces the "semantic equivalence" difficulty in the uncertainty quantification of free-form LLMs and tackles this issue by gathering sentences containing the same meaning into clusters for cluster-wise entropy calculation.

Table 1
Automatic evaluation of open-ended question answering is highly accurate as compared to human evaluation on mPLUG-Owls3. We evaluate how accurate the automatic evaluation metric. The predictions, in this settings are the automatically determined accuracy labels on our open-ended question answering task, and the ground truth are human labels for the accuracy of the provided model generation given the reference answer.

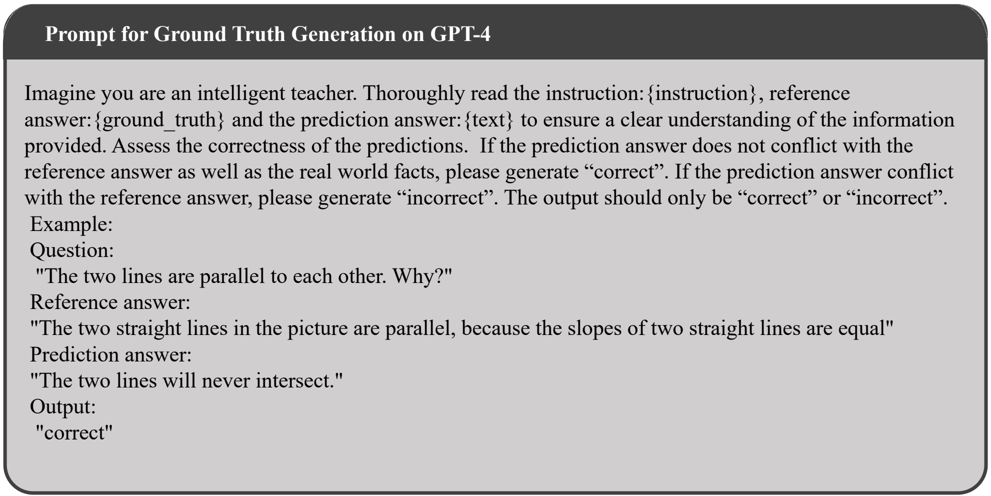
Dataset	Accuracy
Bingo	85.6
MSCOCO-Caption	83.2

LVLMs. We conduct experiments over 5 popular "off-the-shelf" LVLMs, including GPT-4o ("o" for "omni"), LLaMA-3.2-Vision-Instruct-11B, LLaVA-1.5 with model size 7B and 13B, and mPLUG-Owl3-7B.

Benchmarks. For the open-ended benchmarks, We consider 2 open-ended question answering datasets to assess the performance of existing LVLMs: (1) Bingo: This dataset is designed to evaluate and shed light on two common types of hallucinations in LVLMs: bias and interference. Each image in Bingo is paired with one or two questions; (2) MSCOCO-Caption: This dataset is simply designed for image captioning task. Images are sampled from the MSCOCO validation set. Besides, to standardize the output format of each model and ensure the comparison results are authoritative, models are required to output only two sentences uniformly.

Considering that open-ended benchmarks generally have higher variability than closed-ended question, which is hard to get ground truth answers before model inference, we use GPT-4o with specifically designed prompt to annotate the hallucinations, see Figure.6. Besides, we hire some human labelers to annotate the hallucinations as well. They were paid 0.15$ per HIT, which is more than prevailing local minimum wage. Human evaluation verifies that our constructed benchmark based on GPT-4o can adequately evaluate the performance of the hallucination evaluation methods, see Table 1. Such an approach enables completely automatic evaluation and allows us to scale up our experiments.

Implementation Details. For all LVLMs besides GPT-4o, we extract probabilities of top-10 likely output for token. For GPT-4o, we generate top-5 likely output probabilities. The max sequence length of LVLMs output is set to 1024, and the temperature t is set to 0 for our evaluation work. If the last sentence of the generation is unfinished (i.e. does not end with any punctuation), it is discarded. All the others are set as default. All the experiments are conducted on a server



Prompt for Ground Truth Generation on GPT-4

Imagine you are an intelligent teacher. Thoroughly read the instruction:{instruction}, reference answer:{ground_truth} and the prediction answer:{text} to ensure a clear understanding of the information provided. Assess the correctness of the predictions. If the prediction answer does not conflict with the reference answer as well as the real world facts, please generate "correct". If the prediction answer conflict with the reference answer, please generate "incorrect". The output should only be "correct" or "incorrect".
Example:
Question:
"The two lines are parallel to each other. Why?"
Reference answer:
"The two straight lines in the picture are parallel, because the slopes of two straight lines are equal"
Prediction answer:
"The two lines will never intersect."
Output:
"correct"

Figure 6: Prompt for ground truth generation on GPT-4o.

Table 2
Comparison with state-of-the-arts on free-form benchmark (MSCOCO-Cap and Bingo) for LVLMs hallucination detection. In each setting, the bold **value** indicates the best result, while the underlined value represents the sub-optimal. Our Sarah yields significant improvements over strong baselines. This validates the efficacy of our proposed method, which further facilitates LVLMs hallucination detection. The reported results are hallucination detection accuracy. We re-implement semantic entropy and InterrogateLLM within vision-language context. Besides, due to GAVIE's requirement for bounding box ground truth, we conducted GAVIE only on the MSCOCO-CAP dataset that provides this part of the data.

Models & Datasets	GAVIE	FAITHSCORE	Semantic Entropy	InterrogateLLM	HalDet-LLaVA	VL-Uncertainty	Sarah
LLaVA-v1.5-7B							
Bingo		63.6	62.4	**66.0**	60.3	58.0	64.4
MSCOCO-Cap	56.6	61.6	54.7	**71.7**	50.2	2.7	71.3
LLaVA-v1.5-13B							
Bingo		59.5	64.7	66.8	57.2	60.4	**70.7**
MSCOCO-Cap	67.9	64.2	53.7	72.0	52.7	2.5	**72.6**
GPT-4o							
Bingo		50.8	55.2	**63.2**	37.7	59.3	61.6
MSCOCO-Cap	71.6	69.8	54.6	**86.6**	34.6	4.2	**86.6**
mPLUG-Owl3-7B							
Bingo		60.3	60.6	**71.4**	54.3	48.1	65.1
MSCOCO-Cap	63.1	60.6	53.9	**77.5**	47.0	1.7	70.5
Llama-3.2-Vision-11B							
Bingo		54.2	55.5	**67.7**	49.6	60.0	62.4
MSCOCO-Cap	59.7	68.7	52.5	**78.8**	31.8	2.5	71.2

with one Intel(R) Xeon(R) Platinum 8352V CPU and four NVIDIA 4090 GPUs (python version: 3.10.8) .

4.1. Sarah for Hallucination Detection of LVLMs

Comparison with State-of-the-Art Methods. We evaluate Sarah against several state-of-the-art (SOTA) hallucination detection methods on open-ended benchmarks, with results summarized in Table 2. The detection performance of our method is competitive with InterrogateLLM and significantly outperforms other comparative methods. In direct comparison with InterrogateLLM, Sarah surpasses InterrogateLLM in three scenarios while falling slightly behind (by no more than 1.6%) in three cases. This marginal disadvantage is considered acceptable. Notably, although Sarah exhibits slightly inferior detection accuracy compared to InterrogateLLM, it demonstrates substantial advantages in both time and computational resource consumption. As shown in Table 3, Sarah requires only 1/25 of the detection time needed by InterrogateLLM. This efficiency stems from Sarah's independence from extensive external tools and its single-round inference requirement. Furthermore, the unique hallucination quantification method through uncertainty calculation eliminates the need for re-reading input

Table 3
Resource consumption across diverse hallucination detection methods. This analysis evaluates whether a detection method requires (1) multi-round reasoning and (2) image inputs, both of which significantly affect GPU memory requirements. The actual iteration time for each method is also provided in the table.

Method	Multi-round	Input image	Time
SE			3.65
Faithscore		*	6.02s/it
GAVIE		*	5.48s/it
InterrogateLLM	*		109.56 s/it
HalDet-LLaVA		*	4.35 s/it
VL-Uncertainty	*	*	27.96/it
Sarah			4.48

images for analysis, thereby significantly reducing GPU memory demands.

Our method demonstrates substantially more pronounced advantages compared to other hallucination detection methods across all evaluated LVLMs with varied architectures and sizes. We observed that Sarah achieves performance gains of +9. 7% for LLaVA-V1.5-7B, +4. 7% for LLaVA-V1.5-13B, 15. 0% for GPT-4o, +5. 4% for mPLUG-Owl3-7B and +4. 5% for Llama-3.2-Vision-11B in MSCOCO-Cap. The robustness of Sarah is further validated on the Bingo dataset, where it achieves improvements of +0. 8% for LLaVA-V1.5-7B, +6. 0% for LLaVA-V1.5-13B, +2. 3% for GPT-4o, +4. 5% for mPLUG-Owl3-7B and +2. 4% for Llama-3.2-Vision-11B.

These consistent advancements over strong baselines benefit from Sarah's comprehensive and rational allocation of attention to the output content, coupled with thorough exploitation of the output probability distribution data. As illustrated in Figure.7, while both Sarah and SE support token-level evaluation, Sarah demonstrates superior performance in accurately identifying hallucinatory tokens instead of high uncertainty token merely. GAVIE exhibits object-centric limitations: while effective for noun-related hallucinations, it fails to detect hallucinatory verbs or other non-nominal syntactic elements. Other methods, though capable of detecting hallucinations in this context, operates solely at the sentence-level granularity, resulting in inevitable detail loss.

In the MSCOCO-Captioning task, VL-Uncertainty demonstrates accuracy failing to exceed 2.7% under all experimental conditions. This limitation primarily stems from its evaluating hallucination severity by perturbing input multiple times and computing the uncertainty of outputs. However, in open-ended tasks like image captioning, valid outputs can vary significantly in wording while remaining semantically correct, making uncertainty estimation unreliable. Thus, while VL-Uncertainty works reasonably on structured tasks, it fails in open-ended generation.

4.2. Hallucination Performance of Different LVLMs

As shown in Table 4, we present a comprehensive performance comparison of various models in terms of Sarah when benchmarked on the Bingo and MSCOCO-Cap datasets. We have the following observations: (1) GPT-4o outperforms other counterparts in most situations. This demonstrates its preeminent capability in generating content that is factually accurate and consistent with the input. (2) For most models, the performance on the MSCOCO-Cap dataset is better than their performance on the Bingo dataset. The potential reason may be that the MSCOCO-Cap dataset only requires the model to master the ability to describe image content, while Bingo requires the model to further overcome the bias caused by the imbalance in training data, as well as the interference caused by the format of the input text prompts or input images. (3) It's worth noting that besides LLaVA-V1.5, different models have similar performance distribution across tasks. They are all better at OCR and factual reasoning tasks, but have poor adversarial robustness against interference in the input prompts and images. For instance, Llama-3.2-Vision-11b achieved 0.4574, 0.3563, 0.4667, 0.5446 and 0.5605 on the "Region Bias", "OCR Bias", "Factual Bias", "Image-to-Image Interference" and "Text-to-Image Interference" tasks, respectively. (4) Compared to others, LLaVA-V1.5 performs worse in most tasks and does not evidently excel in any specific task.

4.3. Ablation Studies

To evaluate the contributions of each proposed modules to the effectiveness of Sarah, we conducted an ablation study focusing on the Semantic Information Locator and Semantic Information Purifier modules. Additionally, we used average token entropy as a baseline for comparison. The experiments were performed on the Bingo dataset using GPT-4o as the generator. The results are summarized in Table 5, from which we derive the following key observations.

Only Semantic Information Locator. The introduction of the Semantic Information Locator significantly improved hallucination detection accuracy by 25. 3%, elevating the model from near-random guessing performance (30. 9%) to a robust tool capable of effective hallucination evaluation (56. 2%). This demonstrates that the locator accurately identifies hallucinated content by leveraging the uneven distribution of semantic importance across tokens. The results align with our hypothesis that semantic importance weighting is critical for precise hallucination localization.

Only Semantic Information Purifier. Similarly, the Semantic Information Purifier demonstrates strong performance in hallucination detection, achieving an accuracy higher than the baseline (35.8%). This highlights the module's ability to refine uncertainty quantification by filtering out expression-level variations and focusing on semantic content discrepancies. The purifier's superior performance underscores its role in reducing noise and enhancing the reliability of hallucination detection.

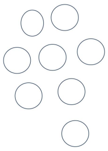
Dataset	Image	Prompt	Method	Attention of Different Methods on Hallucinations
MSCOCO-Cap		Generate a short caption of the image.	Sarah	Someone is checking something on a device.
			SE	Someone is checking something on a device.
			GAVIE	Someone is checking something on a device.
			FaithScore	Someone is checking something on a device.
			InterrogateLLM	Someone is checking something on a device.
			HalDet-LLaVA	Someone is checking something on a device.
			VL-Uncertainty	Someone is checking something on a device.
Bingo		One of the shapes is an ellipse. What do you think?	Sarah	None of the shapes are ellipses.
			SE	None of the shapes are ellipses.
			FaithScore	None of the shapes are ellipses.
			InterrogateLLM	None of the shapes are ellipses.
			HalDet-LLaVA	None of the shapes are ellipses.
			VL-Uncertainty	None of the shapes are ellipses.

Figure 7: Visualization of attention allocation patterns in hallucination detection for GPT-4 generated text (analyzed by Sarah, GAVIE, SE, InterrogateLLM, HalDet-LLaVA and VL-Uncertainty). Darker yellow hues indicate higher attention weights

Table 4
Hallucination rate of different LVLMs as evaluated by Sarah.

Model	Bingo						MSCOCO-Cap
	Overall	Region	OCR	Factual	i2i	t2i	
LLaVA-v1.5-7B	61.3	57.9	64.0	61.6	62.9	61.3	29.0
LLaVA-v1.5-13B	59.2	64.2	54.0	54.1	65.0	58.7	26.8
GPT-4o	**38.4**	**39.8**	**16.3**	**36.7**	**44.0**	**49.3**	**13.4**
mPLUG-Owl3-7B	52.0	51.1	48.1	50.7	54.8	54.8	29.2
Llama-3.2-Vision-Instruct-11B	49.0	45.8	35.6	46.7	54.5	56.1	21.0

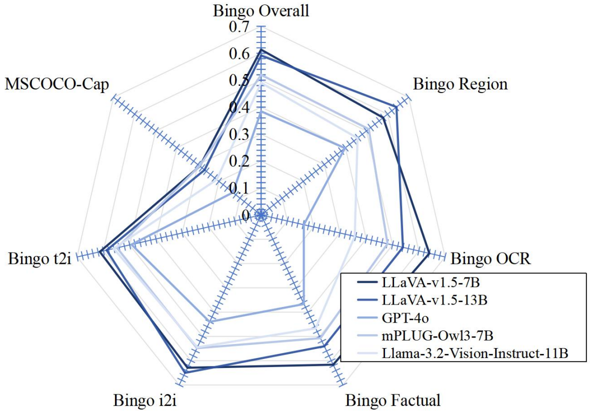


Figure 8: Hallucination rate of different LVLMs based on Sarah.

Fusion Strategy. By adaptively cascading the Semantic Information Locator and Semantic Information Purifier, incorporating them with a multi-threshold decision mechanism, our method achieves a detection accuracy of 86.6%, surpassing the performance of either module in isolation. Notably, the purifier alone outperforms the locator by 10.50%, indicating that the hallucination detection framework relies more heavily on filtering and refining the output content itself rather than solely locating semantic hotspots. This finding emphasizes the importance of carefully selecting adaptive parameters and thresholds to balance the contributions of both modules.

Table 5
Ablation study of contribution of semantic information locator and semantic information purifier for hallucination detection.

Method	Accuracy (%)
Baseline	30.9
+ Locator Only	56.2
+ Purifier Only	66.7
Full Model (Sarah)	**86.6**

Length of Output. We further report an ablation study on comparing the influence of length of generation on hallucination detection. Here we categorize the length of the output content into three groups: (0, 20), (20, 40), and over 40.

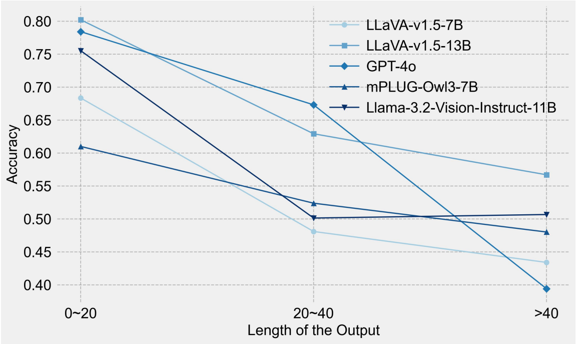


Figure 9: Ablation study of length of output.

As illustrated in Figure.9, the performance of Sarah varies with the length of output. While all models start with relatively high accuracy, their performance drops as the length of the prediction increases. For example, LLaVA-v1.5-7B starts with a high accuracy of 68.35% for length between 0 to 20 and sustains a relatively low accuracy of 43.41% for length over 40.

5. Conclusion

In this work, we propose Sarah, a novel uncertainty quantification based approach for evaluating and detecting hallucination on open-ended question-answering tasks for LVLMs. The proposed method requires merely a single inference pass without image inputs and enables fine-grained measurement of hallucination-contributing positions. This capability is achieved through our novel semantic information locator and purifier. Experiments over "off-the-shelf" LVLMs demonstrate the superior performances of Sarah in most tasks, achieving detection accuracy of 86.6% at most. Notably, while Sarah only achieves comparable performance to InterrogateLLM (tying in evaluation metrics), its substantial computational efficiency (requiring only 1/25 of InterrogateLLM's resources) effectively compensates for this performance limitation.Furthermore, the performance exhibits a degradation trend with increasing generation length, suggesting opportunities for feature space optimization. Our quantitive analysis also demonstrates that current LVLMs are prone to hallucination problems. We hope our work can help address the unexpected hallucination issues of LVLMs. Future directions include further eliminating the interference caused by the uncertainty of language style and word order, as well as better integrating different semantic processing techniques to achieve a synergistic effect where 1+1>2.

6. Limitation

Our method evaluates only the single-round inference result of the model, which may entail some degree of randomness. However, it also significantly reduces the demand for computational resources and processing time while maintaining a certain level of accuracy. Besides, the utilization of sentence similarity calculations and semantic judgement might bring additional latency in practice. In addition, our methods require access to token logits. It still might restrict the potential applications of our methods. Our proposed method has the potential to impact the credibility and reliability of LVLMs, particularly in the context of highlighting parts that are highly likely to be hallucinations.

References

Blundell, C., Cornebise, J., Kavukcuoglu, K., Wierstra, D., 2015. Weight uncertainty in neural network, in: International conference on machine learning, PMLR. pp. 1613–1622.

Chen, J., Guo, H., Yi, K., Li, B., Elhoseiny, M., 2022. Visualgpt: Data-efficient adaptation of pretrained language models for image captioning, in: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition, pp. 18030–18040.

Chen, X., Wang, C., Xue, Y., Zhang, N., Yang, X., Li, Q., Shen, Y., Liang, L., Gu, J., Chen, H., 2024. Unified hallucination detection for multimodal large language models, in: ACL (1).

Cui, C., Ma, Y., Cao, X., Ye, W., Zhou, Y., Liang, K., Chen, J., Lu, J., Yang, Z., Liao, K.D., et al., 2024. A survey on multimodal large language models for autonomous driving, in: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision, pp. 958–979.

Cui, C., Zhou, Y., Yang, X., Wu, S., Zhang, L., Zou, J., Yao, H., 2023. Holistic analysis of hallucination in gpt-4v (ision): Bias and interference challenges. arXiv preprint arXiv:2311.03287 .

De Marneffe, M.C., Manning, C.D., Nivre, J., Zeman, D., 2021. Universal dependencies. Computational linguistics 47, 255–308.

Duan, J., Cheng, H., Wang, S., Zavalny, A., Wang, C., Xu, R., Kailkhura, B., Xu, K., 2023. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. arXiv preprint arXiv:2307.01379 .

Fadeeva, E., Rubashevskii, A., Shelmanov, A., Petrakov, S., Li, H., Mubarak, H., Tsymbalov, E., Kuzmin, G., Panchenko, A., Baldwin, T., et al., 2024. Fact-checking the output of large language models via token-level uncertainty quantification. arXiv preprint arXiv:2403.04696 .

Gal, Y., Ghahramani, Z., 2016. Dropout as a bayesian approximation: Representing model uncertainty in deep learning, in: international conference on machine learning, PMLR. pp. 1050–1059.

Hu, H., Zhang, J., Zhao, M., Sun, Z., 2023a. Ciem: Contrastive instruction evaluation method for better instruction tuning. arXiv preprint arXiv:2309.02301 .

Hu, Y., Liu, B., Kasai, J., Wang, Y., Ostendorf, M., Krishna, R., Smith, N.A., 2023b. Tifa: Accurate and interpretable text-to-image faithfulness evaluation with question answering, in: Proceedings of the IEEE/CVF International Conference on Computer Vision, pp. 20406–20417.

Huang, L., Yu, W., Ma, W., Zhong, W., Feng, Z., Wang, H., Chen, Q., Peng, W., Feng, X., Qin, B., et al., 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. ACM Transactions on Information Systems 43, 1–55.

Huang, S., Dong, L., Wang, W., Hao, Y., Singhal, S., Ma, S., Lv, T., Cui, L., Mohammed, O.K., Patra, B., et al., 2023. Language is not all you need: Aligning perception with language models. Advances in Neural Information Processing Systems 36, 72096–72109.

Jin, Y., Li, J., Liu, Y., Gu, T., Wu, K., Jiang, Z., He, M., Zhao, B., Tan, X., Gan, Z., et al., 2024. Efficient multimodal large language models: A survey. arXiv preprint arXiv:2405.10739 .

Jing, L., Li, R., Chen, Y., Du, X., 2023. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. arXiv preprint arXiv:2311.01477 .

Kadavath, S., Conerly, T., Askell, A., Henighan, T., Drain, D., Perez, E., Schiefer, N., Hatfield-Dodds, Z., DasSarma, N., Tran-Johnson, E., et al., 2022. Language models (mostly) know what they know. arXiv preprint arXiv:2207.05221 .

Kendall, A., Gal, Y., 2017. What uncertainties do we need in bayesian deep learning for computer vision? Advances in neural information processing systems 30.

Kuhn, L., Gal, Y., Farquhar, S., 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. arXiv preprint arXiv:2302.09664 .

Leng, S., Xing, Y., Cheng, Z., Zhou, Y., Zhang, H., Li, X., Zhao, D., Lu, S., Miao, C., Bing, L., 2024. The curse of multi-modalities: Evaluating hallucinations of large multimodal models across language, visual, and audio. arXiv preprint arXiv:2410.12787 .

Li, Q., Geng, J., Lyu, C., Zhu, D., Panov, M., Karray, F., 2024. Reference-free hallucination detection for large vision-language models. arXiv preprint arXiv:2408.05767 .

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R., 2023. Evaluating object hallucination in large vision-language models. arXiv preprint arXiv:2305.10355 .

Liang, Z., Xu, Y., Hong, Y., Shang, P., Wang, Q., Fu, Q., Liu, K., 2024. A survey of multimodel large language models, in: Proceedings of the 3rd International Conference on Computer, Artificial Intelligence and Control Engineering, pp. 405–409.

Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L., 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, Springer. pp. 740–755.

Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L., 2023a. Mitigating hallucination in large multi-modal models via robust instruction tuning. arXiv preprint arXiv:2306.14565 .

Liu, F., Lin, K., Li, L., Wang, J., Yacoob, Y., Wang, L., 2023b. Mitigating hallucination in large multi-modal models via robust instruction tuning. arXiv preprint arXiv:2306.14565 .

Liu, H., Xue, W., Chen, Y., Chen, D., Zhao, X., Wang, K., Hou, L., Li, R., Peng, W., 2024. A survey on hallucination in large vision-language models. arXiv preprint arXiv:2402.00253 .

Manakul, P., Liusie, A., Gales, M.J., 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. arXiv preprint arXiv:2303.08896 .

Mao, Y., Lu, D., Zhang, Y., Wang, X., 2023. Fatrer: Full-attention topic regularizer for accurate and robust conversational emotion recognition, in: ECAI 2023. IOS Press, pp. 1688–1695.

Min, S., Krishna, K., Lyu, X., Lewis, M., Yih, W.t., Koh, P.W., Iyyer, M., Zettlemoyer, L., Hajishirzi, H., 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. arXiv preprint arXiv:2305.14251 .

Moor, M., Huang, Q., Wu, S., Yasunaga, M., Dalmia, Y., Leskovec, J., Zakka, C., Reis, E.P., Rajpurkar, P., 2023. Med-flamingo: a multimodal medical few-shot learner, in: Machine Learning for Health (ML4H), PMLR. pp. 353–367.

OpenAI, 2023. GPT-4o: The Next Generation of Language Models with Enhanced Vision Capabilities. https://openai.com/blog/gpt-4o/. Accessed: 2024-05-14.

Park, J.S., O'Brien, J., Cai, C.J., Morris, M.R., Liang, P., Bernstein, M.S., 2023. Generative agents: Interactive simulacra of human behavior, in: Proceedings of the 36th annual acm symposium on user interface software and technology, pp. 1–22.

Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F., 2023. Kosmos-2: Grounding multimodal large language models to the world. arXiv preprint arXiv:2306.14824 .

Rohrbach, A., Hendricks, L.A., Burns, K., Darrell, T., Saenko, K., 2018. Object hallucination in image captioning. arXiv preprint arXiv:1809.02156 .

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., Klimov, O., 2017. Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 .

Stock, A., Schlögl, S., Groth, A., 2023. Tell me, what are you most afraid of? exploring the effects of agent representation on information disclosure in human-chatbot interaction, in: International Conference on Human-Computer Interaction, Springer. pp. 179–191.

Strohauer, S., Wietschorke, F., Zugliani, L., Flaschmann, R., Schmid, C., Grotowski, S., Müller, M., Jonas, B., Althammer, M., Gross, R., et al., 2023. Site-selective enhancement of superconducting nanowire single-photon detectors via local helium ion irradiation. Advanced Quantum Technologies 6, 2300139.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 .

Wang, J., Zhou, Y., Xu, G., Shi, P., Zhao, C., Xu, H., Ye, Q., Yan, M., Zhang, J., Zhu, J., et al., 2023a. Evaluation and analysis of hallucination in large vision-language models. arXiv preprint arXiv:2308.15126 .

Wang, W., Xie, J., Hu, C., Zou, H., Fan, J., Tong, W., Wen, Y., Wu, S., Deng, H., Li, Z., et al., 2023b. Drivemlm: Aligning multi-modal large language models with behavioral planning states for autonomous driving. arXiv preprint arXiv:2312.09245 .

Yehuda, Y., Malkiel, I., Barkan, O., Weill, J., Ronen, R., Koenigstein, N., 2024. Interrogatellm: Zero-resource hallucination detection in llm-generated answers. arXiv preprint arXiv:2403.02889 .

Zhang, G., Gao, H.a., Jiang, Z., Zhao, H., Zheng, Z., 2024a. Ctrl-u: Robust conditional image generation via uncertainty-aware reward modeling. arXiv preprint arXiv:2410.11236 .

Zhang, J., Huang, J., Jin, S., Lu, S., 2024b. Vision-language models for vision tasks: A survey. IEEE Transactions on Pattern Analysis and Machine Intelligence .

Zhang, R., Zhang, H., Zheng, Z., 2024c. Vl-uncertainty: Detecting hallucination in large vision-language model via uncertainty estimation. arXiv preprint arXiv:2411.11919 .

Declaration of interests

☒The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: