Seeing is Believing: Rich-Context Hallucination Detection for MLLMs via Backward Visual Grounding

Pinxue Guo[1*], Chongruo Wu[3*], Xinyu Zhou[2], Lingyi Hong[2], Zhaoyu Chen[1], Jinglun Li[1], Kaixun Jiang[1], Sen-Ching Samson Cheung[4], Wei Zhang[2†], Wenqiang Zhang[1,2†]

[1]College of Intelligent Robotics and Advanced Manufacturing, Fudan University
[2]College of Computational Science and Artificial Intelligence, Fudan University
[3]Independent Researcher
[4]Electrical and Computer Engineering, University of Kentucky

Abstract

Multimodal Large Language Models (MLLMs) have unlocked powerful cross-modal capabilities, but still significantly suffer from hallucinations. As such, accurate detection of hallucinations in MLLMs is imperative for ensuring their reliability in practical applications. To this end, guided by the principle of "Seeing is Believing", we introduce `VBackChecker`, a novel reference-free hallucination detection framework that verifies the consistency of MLLM-generated responses with visual inputs, by leveraging a pixel-level Grounding LLM equipped with reasoning and referring segmentation capabilities. This reference-free framework not only effectively handles rich-context scenarios, but also offers interpretability. To facilitate this, an innovative pipeline is accordingly designed for generating instruction-tuning data (`R-Instruct`), featuring rich-context descriptions, grounding masks, and hard negative samples. We further establish `R`^{2}`-HalBench`, a new hallucination benchmark for MLLMs, which, unlike previous benchmarks, encompasses real-world, rich-context descriptions from 18 MLLMs with high-quality annotations, spanning diverse object-, attribute-, and relationship-level details. `VBackChecker` outperforms prior complex frameworks and achieves state-of-the-art performance on `R`^{2}`-HalBench`, even rivaling GPT-4o's capabilities in hallucination detection. It also surpasses prior methods in the pixel-level grounding task, achieving over a 10% improvement.

Code and Data —
https://github.com/PinxueGuo/VBackChecker

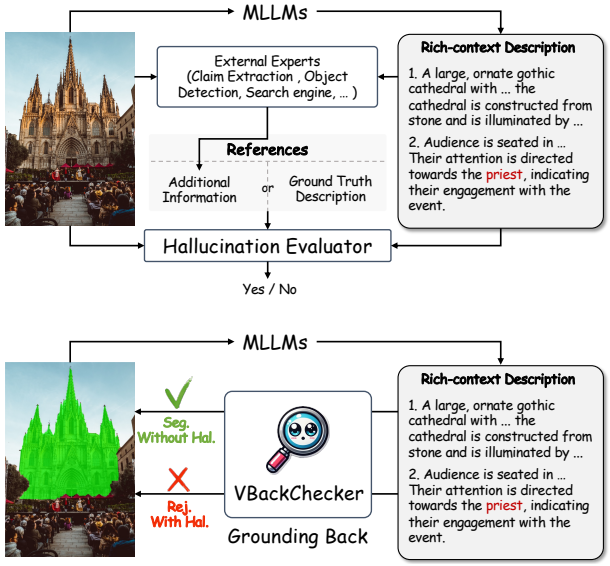

Figure 1: Illustration of our proposed `VBackChecker`. Unlike previous methods (Top), our `VBackChecker` (Bottom) identifies hallucination in MLLMs by leveraging a pixel-level grounding approach for backward verification, without any references or external experts.

Introduction

Multimodal Large Language Models (MLLMs) (Zhu et al. 2023; Liu et al. 2023c; Alayrac et al. 2022; Li et al. 2024; Wang et al. 2024a; Chen et al. 2024c; Panagopoulou et al. 2024) have recently shown strong cross-modal understanding and reasoning capabilities, excelling in image captioning, visual question answering, and multimodal reasoning. However, their effectiveness is still substantially limited by visual hallucination (Liu et al. 2024; Bai et al. 2024), where generated content diverges from the visual input. As hallucination severely undermines reliability and deployment, detecting hallucinations in MLLM-generated rich-context responses is increasingly critical.

Existing methods fall into reference-based and reference-free approaches. Reference-based methods rely on groundtruth text or external expert models (e.g., dense captioning, object detection) to compare with generated content (Wang et al. 2023c; Liu et al. 2023a; Wang et al. 2023a; Chen et al. 2024a). Their dependence on annotated GT or expert models restricts scalability to unannotated inputs and inherits capability limits from these experts. Reference-free methods are more flexible; FaithScore (Jing et al. 2023) leverages VQA to identify hallucinations without references, but its binary responses are biased and lack interpretability. Further, its reliance on LLM-based decomposition into atomic units introduces computation

[*]Equal contribution.
[†]Corresponding authors.

Evaluation Method	Reference Free	Without External Expert	Without Calling GPT API	Explainability		End2End Single Model
				Visual	Language	
HaELM		✓				✓
AMBER			✓			
UniHD					✓	
GAVIE					✓	
FaithScore	✓	✓				
VBackChecker	✓	✓	✓	✓	✓	✓

Benchmark	Hallucination Type	Source MLLMs	Real Dis	Rich-Context Description	Description Avg Len	Max Len
POPE	Obj	0			2.2	3
CIEM	Obj, Attr	1			-	-
AMBER	Obj, Attr, Rel	0			2.4	9
MHaluBench	Obj, Attr, Text	5		✓	13.0	58
R^2-HalBench	Obj, Attr, Rel	18	✓	✓	16.4	110

Table 1: **(Left) Comparison of Hallucination Detection Methods**. **(Right) Comparison of Hallucination Detection Benchmarks**. "Obj, Attr and Rel" represent abbreviations for Object, Attribute and Relation, respectively. "Source MLLMs" denotes the number of MLLMs employed to generate response, with "Real Dis" indicating whether responses reflect the real distribution. The "Description Avg/Max len" represents the avg/max response lengths, measured in word count.

overhead and error propagation, limiting its ability to process rich-context descriptions.

We introduce Verification by Vision Back (VBackChecker), a novel framework (Fig. 1) that addresses these challenges by combining pixel-level grounding (e.g., LISA (Lai et al. 2023)) with reasoning and referring segmentation. The core principle is simple: Seeing is Believing—if described elements can be grounded in the image, no hallucination exists; otherwise, hallucinations are detected. As shown in Table 1 (Left), VBackChecker is fully reference-free, supports arbitrary inputs, and provides interpretable outputs in both vision and language, without external experts.

To enable VBackChecker to handle rich-context scenarios, we propose two key innovations. First, we develop an automatic pipeline to construct instruction-following data, R-Instruct, emphasizing rich-context descriptions paired with grounding masks and hard negative samples. Second, we enhance learning of special tokens [SEG]/[REJ] to mitigate inconsistencies between autoregressive training and our objective. With these designs, VBackChecker achieves strong hallucination detection while improving pixel-level grounding performance by over 10

To evaluate real-world hallucination detection, we propose R^2-HalBench, a Real-response, Rich-context Hallucination Benchmark. As shown in Table 1 (Right), compared with prior benchmarks (Li et al. 2023; Hu et al. 2023; Wang et al. 2023b; Chen et al. 2024a), R^2-HalBench provides superior coverage: real responses from 18 advanced MLLMs across diverse architectures, high-quality human annotations, and significantly longer and richer descriptions containing object-, attribute-, and relationship-level details. This makes it suitable for assessing both hallucination detectors and MLLM hallucination behaviors.

In summary, our contributions are:

- We introduce VBackChecker, the first visual-back grounding framework for hallucination detection in MLLMs, functioning reference-free, supporting rich-context responses, and offering interpretability across modalities.
- We develop R-Instruct, a new instruction-tuning dataset focused on rich-context object descriptions with grounding masks and challenging negative samples.
- We propose R^2-HalBench, a real-response, rich-context hallucination detection benchmark with high-quality annotations reflecting real-world hallucination distributions.
- VBackChecker achieves state-of-the-art performance on R^2-HalBench and POPE, rivaling GPT-4o in hallucination detection, while improving pixel-level grounding by over 10

Related Work

Grounding LLM

While large language models exhibit strong reasoning capabilities in NLP tasks, researchers are progressively extending these abilities to the domain of multimodal large language models. Through aligning visual and textual inputs, models (Liu et al. 2023d,b; Zhu et al. 2023) are able to understand visual information, significantly broadening their range of applications. Furthermore, grounding capabilities have also been explored in the multimodal large language models, enabling them not only to comprehend an entire image, but also to understand localized information, or even output precise spatial information (Peng et al. 2023; Chen et al. 2023; Zhang et al. 2023; Lai et al. 2024). Among these, LISA (Lai et al. 2024) was the first work to touch the reasoning segmentation task, which was later extended to multi-object grounding (Yang et al. 2023; Rasheed et al. 2024), multi-round conversation (Wang et al. 2024b) and further introduce rejection capabilities (Xia et al. 2024). However, despite these advances, their ability to accurately comprehend and reject rich-context queries remains limited.

Hallucination Detection

Research on hallucination detection in MLLMs (Bai et al. 2024; Liu et al. 2024) has gained significant attention due to the critical need for reliable multimodal systems. Current approaches can be categorized into reference-based and reference-free methods. The former compares generated content against ground-truth to identify discrepancies (Wang et al. 2023c), or utilizes external expert models to extract visual information that serves as reference points (Liu et al. 2023a; Wang et al. 2023a; Chen et al. 2024a). These approaches face significant limitations in real-world applications due to their dependence on annotated ground truth

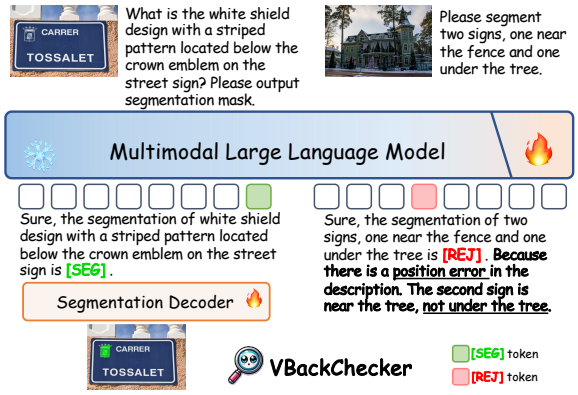

Figure 2: The framework of our proposed VBackChecker

data and external expert models. For reference-free method, FaithScore (Jing et al. 2023) leverages multimodal question-answering capabilities for reference-free hallucination detection. However, the method's binary response paradigm restricts interpretability and renders the approach susceptible to bias. Moreover, the requirement for external experts to break down inputs into basic units adds computational cost and hampers performance on rich-context inputs. In contrast, our proposed method is able to efficiently handle rich-context queries while providing interpretability.

Method

Verification by Visual Back Framework

Problem Formulation. Our work addresses the problem of hallucination detection in Multimodal Large Language Models (MLLMs), which determines whether a given response from MLLMs contains visual hallucinations when conditioned on an input image. Specifically, let an MLLM Φ generate a descriptive response $R = \{r_1, r_2, \ldots, r_n\}$ for an image I, where each rich-context sentence r_i describes a specific visual part of the image: $R = \Phi(I)$. Sentences r_i may include visual hallucinations that conflict with the actual visual input, as the MLLMs rely on learned patterns rather than adhering to the visual information in I. The goal of hallucination detection is to evaluate each r_i and determine whether it contains a hallucination. This is formulated as a binary classification task for each r_i, where the label $y_i \in \{0, 1\}$ indicates the presence or absence of hallucinations. Specifically:

$$y_i = \begin{cases} 0 & \text{if } r_i \text{ aligns with the visual input } I, \\ 1 & \text{if } r_i \text{ conflicts with the visual input } I. \end{cases} \quad (1)$$

VBackChecker Framework. We propose the Verification by Vision Back Framework (VBackChecker), as illustrated in Fig. 1, to tackle this problem by leveraging a pixel-level Grounding LLM with reasoning and referring segmentation capabilities. The key idea behind VBackChecker is straightforward yet logical: Seeing is Believing. It verifies the response by "looking back" at the visual input. For each r_i in the MLLM's generated response, VBackChecker determines whether r_i could be accurately grounded back to the input image I. If the grounding is successful, the sentence is free of hallucinations ($y_i = 0$); otherwise, it indicates the sentence contains a hallucination ($y_i = 1$).

VBackChecker employs a conversational model to predict either a [SEG] token or a [REJ] token to perform hallucination detection: When there is no hallucination in the MLLM's response r_i based on the input image I, VBackChecker outputs a [SEG] token. The embedding of [SEG] before the LLM head is passed to a mask decoder (as in LISA (Lai et al. 2023)) to decode an object mask $m \in \mathbb{R}^{H \times W}$ corresponding to the described part in r_i. When the response r_i contains a hallucination, VBackChecker outputs a [REJ] token and generates a detailed explanation in natural language indicating where r_i conflicts with the visual content of I. By combining pixel-level grounding and language-based reasoning, VBackChecker not only detects hallucinations but also provides interpretable outputs in both visual and language modalities. This dual capability ensures robust and reliable hallucination checking for MLLMs.

However, existing Grounding LLMs lack sufficient discriminative power for rich-context queries. While GSVA (Xia et al. 2024) attempts to reject references to non-existent objects, it struggles with fine-grained understanding. For example, when querying about a "... person in white" in an image with a person in black near a white wall, GSVA incorrectly predicts [SEG] instead of [REJ], failing to distinguish subtle attributes like object-specific colors.

R-Instruct Data Generation

To address this gap, we introduce Rich-context Instruct Tuning data (R-Instruct), built by the automated pipeline designed to generate rich-context visual instruction data for grounding and hallucination detection. R-Instruction includes grounding masks for positive queries and hallucination types with explanations for negative queries, enabling interpretability in both visual and language modalities

As illustrated in Figure 3(a), we propose an automated four-step procedure to construct our R-Instruct data. Let $\mathcal{I}$ be the set of images sampled from the large-scale and high-quality SA1B dataset. For each image $I \in \mathcal{I}$:

Object Proposal. We first employ the Recognize Anything Model (Zhang et al. 2024) and the Grounded Segment Anything Model (Ren et al. 2024) to identify candidate objects $\{\Omega_m\}_{m=1}^{M}$, where each Ω_m includes a bounding box and the corresponding segmentation mask.

Rich-context Caption. For each proposal Ω_m, we employ Qwen2-VL-72B (Wang et al. 2024a) to generate a detailed caption C_m, capturing the category, shape, attributes (e.g., color, texture), spatial relationships, and distinctive features. This yields a set of object-description pairs, ensuring unique identification of objects within the image.

Quality Control. To ensure accurate and diverse captions, by using CLIP (Radford et al. 2021) and SBERT (Reimers and Gurevych 2019), we apply a multi-criteria filtering process: (1) Foreground Background Check. Given a proposal Ω_m, we compute CLIP scores for its foreground-masked image I^m_{fg} and background-masked image I^m_{bg}: $S^m_{\text{fg}} = \text{CLIP}(I^m_{\text{fg}}, C_m)$, $S^m_{\text{bg}} = \text{CLIP}(I^m_{\text{bg}}, C_m)$. We retain captions where $S^m_{\text{fg}} > S^m_{\text{bg}}$, ensuring the description aligns with

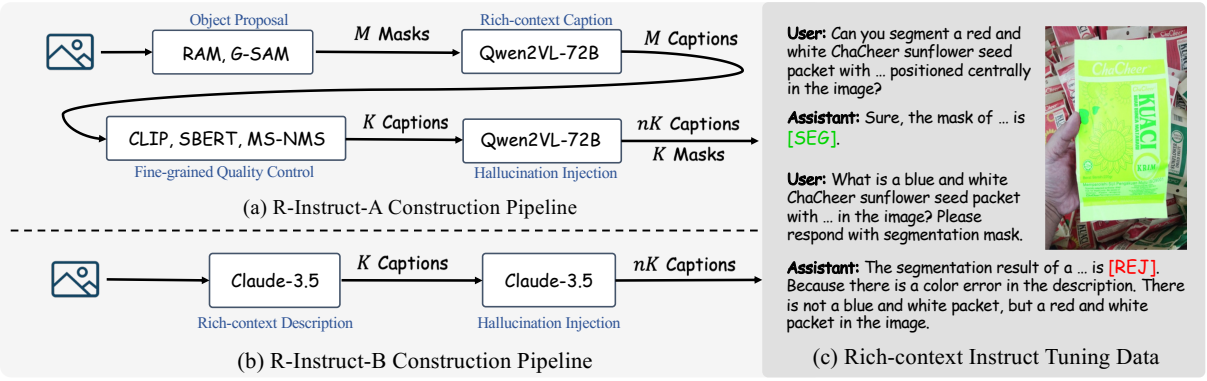

Figure 3: The Generation Pipeline of Rich-Context Instruct Tuning Data (`R-Instruct`).

the object rather than background artifacts. (2) Vision-Text Consistency. Captions with S_{fg}^m below a confidence threshold (0.5) are discarded, filtering out poorly aligned descriptions. (3) Multimodal Semantic NMS (Non-Maximum Suppression). We propose a Multimodal Semantic NMS to remove near-duplicate proposals. We treat textual similarity $\text{sim}(C_m, C_n)$ (computed via SBERT) as "semantic IoU" and use S_{fg}^m as a confidence score. Captions with high similarity to a higher-scoring reference are iteratively removed, ensuring each object retains its most coherent description. This process yields K high-quality samples (Ω_k, C_k), each with a precise mask and a verified rich-context caption.

Hallucination Injection. To create challenging negatives for hallucination detection, we prompt Qwen2-VL to perturb each valid caption C_k using the original image and contextual object descriptions. The model introduces misleading details—e.g., incorrect attributes or nonexistent objects—yielding hallucinated variants. Each is annotated with its hallucination type (e.g., incorrect color, nonexistent object) and an explanation, enabling the model to distinguish accurate from misleading content.

As shown in Figure 3(b), we also collect holistic, multi-object captions (without per-object grounding masks) to better mimic real MLLM outputs. Using Claude-3.5, we generate comprehensive descriptions covering all objects in an image as additional positive samples, then apply the same perturbation process to produce negatives. The resulting dataset(`R-Instruct`) includes 30k images, 300k positive samples (100k with masks), and 500k negatives. Data with masks (Figure 3(a)) forms R-Instruct-A; data without masks (Figure 3(b)) forms R-Instruct-B. Positive responses contain [SEG], while negative ones include [REJ] with hallucination explanations (Figure 3(c)).

Instruction Tuning

The training methodology for Rich-context Instruction Tuning leverages a self-regressive learning framework tailored for multimodal grounding and hallucination detection. The overall loss function consists of two components: language loss ($\mathcal{L}_L$) and grounding loss ($\mathcal{L}_G$), defined as:

$$\mathcal{L} = \mathcal{L}_L(t, t') + \mathcal{L}_G(m, m'), \tag{2}$$

where t and m denote the predicted response and predicted mask, respectively, and t' and m' represent their corresponding ground truths. Specifically, $t = L_{\text{dec}}(h)$, the textual output decoded by the language decoder L_{dec} from the hidden state h. $m = G_{\text{dec}}(\text{MLP}(h[\text{SEG}]), I)$, the predicted grounding mask generated by decoding the vision-guided hidden state $h[\text{SEG}]$, processed through an MLP projector and conditioned on the image I. G_{dec} is the Mask Decoder, whose architecture follows that of SAM (Kirillov et al. 2023).

The language loss $\mathcal{L}_L$ focuses solely on the response portion of the output sequence, evaluating the next-token prediction with Cross-Entropy Loss. The grounding loss $\mathcal{L}_G$ employs a combination of per-pixel Binary Cross-Entropy Loss and DICE Loss to optimize the segmentation mask prediction. Furthermore, unlike standard MLLM instruction tuning, where every output token equally contributes to the supervision signal, the performance of our grounding LLM, `VBackChecker`, hinges on its ability to make accurate predictions for the special tokens [SEG] and [REJ]. These tokens represent the model's decision to either ground or reject a query and are crucial for hallucination detection. To emphasize their importance during training, we enhance their contributions to the loss function by applying higher weights to their Cross-Entropy terms. The modified language loss for these tokens is defined as:

$$\mathcal{L}_L' = -\sum \alpha_i \log P(t_i | t_{<i}), \alpha_i = \begin{cases} \lambda, & \text{if } t_i \in \{\text{[SEG]}, \text{[REJ]}\} \\ 1, & \text{otherwise} \end{cases} \tag{3}$$

where i indexes tokens in the response. $\lambda > 1$ is a hyperparameter that amplifies the contributions of [SEG] and [REJ], guiding the model to emphasize learning critical decisions.

R²-HalBench

To support reliable evaluation of hallucination detection in rich-context MLLM responses, we introduce R²-HalBench, a benchmark with 3,000 rich-context object descriptions, generated by 18 advanced MLLMs, from 10,000 images. Designed to reflect real-world hallucination distributions, it ensures a comprehensive and representative assessment. Figure 4 showcases several samples of the benchmark.

Model	Param	Model	Param
GPT-4o	–	InternVL2.5-8B	8B
Gemini-1.5	–	Cambrian-8B	8B
Claude-3.5	–	Fuyu-8B	8B
InternVL-78B	78B	Qwen2-VL-7B	7B
Qwen2-72B	72B	InstructBLIP	7B
LLaVA-OV-72B	72B	LLaVA-OV-7B	7B
InternVL-38B	38B	InternVL-2B	2B
InternVL-26B	26B	Qwen2-VL-2B	2B
LLaVA-13B	13B	LLaVA-OV-0.5B	0.5B

Dataset	1–5	6–10	11–15	16–20	21–25	26–30	31–40	>40
POPE	100.0%	–	–	–	–	–	–	–
AMBER	100.0%	–	–	–	–	–	–	–
MHalu	3.1%	35.9%	32.2%	19.5%	6.1%	2.1%	–	–
R^2-Hal	3.6%	17.5%	32.9%	24.4%	11.6%	4.9%	3.3%	1.8%

	Object	Attribute	Relation
Ratio	61.2%	24.0%	14.8%

Table 2: A comprehensive analysis of our R^2-HalBench. **(Left)** Model source diversity sorted by parameter size. **(Right-Top)** Context length distribution across datasets. **(Right-Bottom)** Distribution of hallucination types.

Figure 4: Samples of our R^2-HalBench. Each rich-context response, generated from various advanced MLLMs, is systematically annotated with corresponding hallucination categories.

Dataset Collection and Annotation

R²-HalBench consists entirely of real MLLM outputs, preserving natural hallucination patterns rather than artificially injecting errors by LLMs. We prompt MLLMs to generate rich-context descriptions for objects in high-quality images from the SA1B dataset (Kirillov et al. 2023), ensuring alignment with real-world hallucination distributions. Human annotators then verify each description against the corresponding image, classifying samples as "without hallucination" or "with hallucination". Hallucinated cases are further categorized into three types: Object-level (category/existence), Attribute-level (e.g., color, shape, material), and Relation-level (spatial positioning, interactions). Each sample is reviewed by three independent annotators with at least college-level education, and final labels are determined by majority voting. This rigorous process ensures a high-quality benchmark that accurately reflects real-world hallucination phenomena.

Dataset Statistics

R^2-HalBench is designed to replicate real-world hallucination detection scenarios. Table 2 presents a detailed analysis from four key perspectives:

(a) ***Model Diversity*** To ensure a broad representation of responses across different model architectures and scales, 18 different MLLMs are employed to generate responses, covering both open- and closed-source models, and featuring model sizes ranging from 0.5B to 78B parameters.

(b) ***Rich-Context Description Length*** Different from previous benchmarks such as POPE (Li et al. 2023) and AMBER (Wang et al. 2023b), which rely on simple object name queries and lack rich-context descriptions, R^2-HalBench contains diverse and contextually rich descriptions of varying lengths. Even compared to MHaluBench which provides reference-based evaluation, R^2-HalBench features longer and more complex descriptions while covering responses from 18 MLLMs—significantly more than 5 MLLMs used in MHaluBench. This reduces evaluation bias and enhances the robustness of hallucination detection assessments.

(c) ***Hallucination Categories*** R^2-HalBench systematically covers object-, attribute-, and relation-level hallucinations, enabling a thorough evaluation of detection methods and a detailed analysis of how different detectors handle real-world hallucination types in practical applications.

(d) ***Lexical Distribution of Rich-Context Descriptions*** Word frequency distribution analysis in R^2-HalBench reveals a dominance of attributes and positional terms, effectively capturing fine-grained contextual details. This makes the benchmark particularly well-suited for evaluating models that need to discern subtle inaccuracies in descriptions.

Experiments

In this section, we show comprehensive experiments conducted to evaluate the performance of our proposed

Method	LLM	gRefCOCO									R-Instruct-A Val		
		Validation Set			Test Set A			Test Set B			Validation Set		
		IoU	N-acc	T-acc	IoU	N-acc	T-acc	IoU	N-acc	T-acc	IoU	N-acc	T-acc
ReLA (Liu, Ding, and Jiang 2023)	-	63.6	56.4	96.3	70.0	59.0	97.7	61.0	59.9	95.4	-	-	-
LISA-7B (Lai et al. 2023)	Vicuna-7B	61.6	54.7	-	66.3	50.0	-	58.4	51.9	-	-	-	-
LISA-13B (Lai et al. 2023)	Llama2-13B	63.5	55.3	-	68.2	52.2	-	61.8	56.2	-	-	-	-
SAM4MLLM (Chen et al. 2024b)	Llama3-8B	71.9	66.1	-	74.2	63.9	-	65.3	60.0	-	-	-	-
GSVA-7B (Xia et al. 2024)	Vicuna-7B	66.5	62.4	-	71.1	65.3	-	62.2	60.6	-	13.2	0.4	99.8
GSVA-13B (Xia et al. 2024)	Llama2-13B	70.0	66.0	-	73.3	64.7	-	65.5	62.5	-	12.7	0.3	99.8
Ours	Vicuna-7B	80.2	81.3	95.2	74.0	74.3	94.6	70.8	72.8	97.3	63.8	75.3	70.7

Table 3: Performance of Pixel-level Grouding task on gRefCOCO and R-Instruct-A Val. Metrics include IOU, negative accuracy (N-Acc) and true positive accuracy (T-Acc). Our VBackChecker surpasses prior methods in pixel-level grounding task.

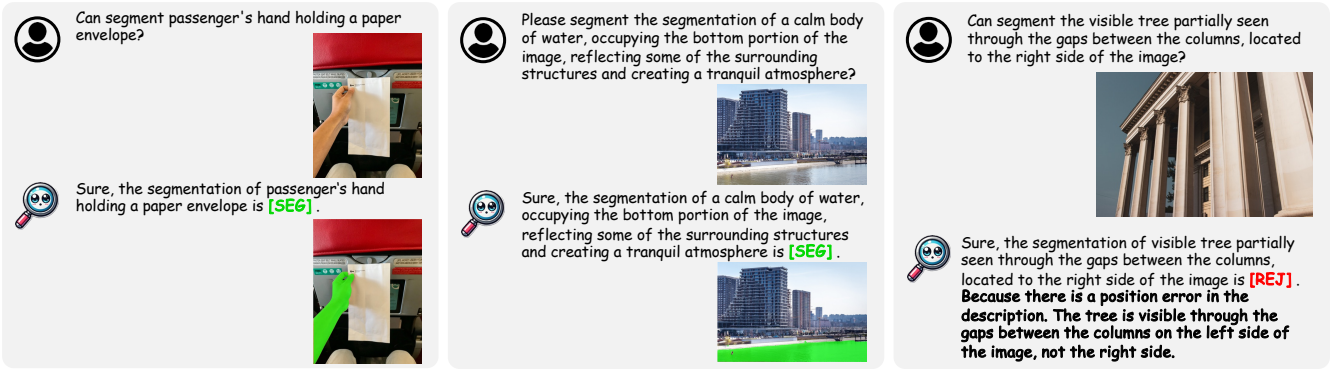

Figure 5: Visualization results of our VBackChecker which employs a conversational model to predict either a [SEG] token or a [REJ] token to perform hallucination detection. Having a [SEG] token indicates that there is no hallucination, otherwise VBackChecker outputs a [REJ] token with detailed explanation.

VBackChecker on both grounding and hallucination detection tasks, as well as ablation studies to analyze the effectiveness and robustness of our framework. The VBackChecker follows LISA in using LLaVa-vicuna-v1.1 and SAM as the base model. It is trained in two stages: the first stage leverages LISA's mixed training dataset along with gRefCOCO, and the second stage uses our proposed R-Instruct data.

Main Results

Pixel-Level Grounding For pixel-level grounding task, we evaluate VBackChecker on two referring segmentation datasets: gRefCOCO (Liu, Ding, and Jiang 2023) and R-Instruct-A-Val, the validation set of R-Instruct-A. Following the baseline GSVA, we adopt three metrics: IoU, negative accuracy (N-Acc) which measures rejection accuracy (correctly rejected queries over groundtruth negative samples), and true positive accuracy (T-Acc) which evaluates segmentation accuracy (correctly segmented queries over groundtruth positive samples). We evaluate VBackChecker on gRefCOCO, which primarily contains simple object references, and on a validation set sampled from the rich-context instruction dataset (Sec) to evaluate its ability to handle more complex queries. VBackChecker shows significant improvements over the previous SOTA method in IoU and rejection accuracy on gRefCOCO (Table 3), achieving 95.2% segmentation accuracy and 81.3% rejection accuracy, demonstrating its effectiveness as a hallucination checker. Additionally, evaluation on R-Instruct-A-Val, a benchmark designed for rich-context queries, shows VBackChecker's superior performance compared to GSVA, which struggles to distinguish hallucinated inputs. Despite the difficulty of benchmark, VBackChecker achieves 75.3% rejection accuracy, confirming its reliability as a hallucination detector.

Hallucination Detection For hallucination detection, VBackChecker is evaluated on the proposed R^2-HalBench which closely mirrors real-world distributions. Figure 5 also shows the detailed visualization results of our framework. The overall accuracy is utilized as the primary metric, along with negative and positive accuracy. In addition, results across different hallucination types, including object-, attribute-, and relation-level hallucinations, are presented in Table 4 to enable a more detailed analysis.

We compare VBackChecker with several reference-free baselines, including reference-free hallucination checkers like FaithScore, grounding-based methods like GSVA, and advanced MLLMs (both open-source and closed-source) that predict hallucination presence directly from an image and query. To address MLLMs' hallucination issues, we design two experimental settings: one where the model explains its decision and the other that provides a binary hallucination prediction.

Despite having only 7B parameters, VBackChecker outperforms all open-source baselines, including FaithScore which employs a more complex pipeline with LLaMA-7B,

Evaluator	LLM	Explainability		R²-HalBench						POPE Acc		
		Visual	Text	Acc	Neg.Acc	Object	Attribute	Relation	Pos.Acc	Random	Popular	Adversarial
GPT-4o (Achiam et al. 2023)	-			63.9	31.4	37.5	23.9	18.6	93.9	96.6	87.4	85.0
GPT-4o (Achiam et al. 2023)	-		✓	68.3	53.8	61.6	43.2	39.2	81.6	98.8	91.4	79.9
LLaVA-OneVision (Li et al. 2024)	7B			47.9	98.8	98.9	98.8	98.5	0.9	93.0	85.6	36.6
LLaVA-OneVision (Li et al. 2024)	7B		✓	48.0	99.1	98.5	99.9	99.9	0.9	99.2	89.1	34.9
Qwen2VL (Wang et al. 2024a)	7B			47.7	76.4	75.6	77.3	78.4	21.2	71.9	68.2	42.7
Qwen2VL (Wang et al. 2024a)	7B		✓	48.9	79.8	80.1	79.2	79.9	20.5	77.5	72.6	41.4
Qwen2VL (Wang et al. 2024a)	72B			48.5	83.2	83.1	86.4	78.4	16.5	64.6	62.3	44.7
Qwen2VL (Wang et al. 2024a)	72B		✓	47.9	81.7	82.4	79.8	81.9	16.8	68.3	65.3	42.7
GSVA (Xia et al. 2024)	7B	✓		49.9	0.5	0.7	0.3	0.0	99.3	23.3	48.3	72.0
GSVA (Xia et al. 2024)	13B	✓		49.8	0.2	0.3	0.0	0.0	99.5	0.5	10.8	64.9
FaithScore	7B+13B+GPT-3.5			59.3	75.3	76.7	73.7	71.5	44.6	82.7	83.0	78.9
Ours	7B	✓	✓	62.5	64.1	64.2	63.4	64.9	60.9	99.3	93.2	70.3

Table 4: Performance of Hallucination Detection on R^2-HalBench and POPE. By leveraging the backward grounding, VBackChecker achieves superior performance and offers interpretability on both modalities, despite of only 7B parameters.

VLM Source	Acc	Neg.Acc	Pos.Acc	VLM Source	Acc	Neg.Acc	Pos.Acc
Claude-3-5-Sonnet	63.2	38.3	71.3	InternVL-2.5-2B	61.6	77.3	45.8
GPT-4o	54.1	56.3	52.5	InternVL-2.5-8B	60.2	72.7	45.8
Gemini-1.5-Pro	56.2	59.6	54.7	InternVL-2.5-26B	63.4	64.8	62.1
Qwen2-VL-2B	62.8	52.9	70.5	InternVL-2.5-38B	65.1	76.2	55.6
Qwen2-VL-7B	67.9	62.3	75.9	InternVL-2.5-78B	61.5	76.1	46.0
Qwen2-VL-72B	59.5	50.8	69.6	LLaVA-OneVision-0.5B	59.7	55.9	64.3
InstructBLIP-7B	69.4	55.0	79.3	LLaVA-OneVision-7B	68.4	66.0	69.8
Cambrain-8B	57.3	56.8	58.2	LLaVA-OneVision-72B	63.5	53.3	72.6
Fuyu-8B	73.7	73.1	75.0	LLaVa-1.5-13B	46.2	31.8	64.7

Context Len (words)	Acc	Neg.Acc	Pos.Acc
1-5	67.3	56.5	75.9
6-10	61.9	54.2	68.3
11-15	61.8	63.3	60.4
16-20	59.1	64.4	54.6
21-30	60.1	68.4	50.9
31-50	63.7	67.2	59.7
51-110	59.1	61.5	55.6

Table 5: **(Left)** VBackChecker's evaluation results of samples from different MLLMs in our R^2-HalBench. **(Right)** Evaluation results of samples with various different length, measured in word count. Our approach maintains consistency and reliability.

Model	R²-HalBench			POPE		
	Acc	NegAcc	PosAcc	Random	Popular	Adv
Ours (VBackChecker)	62.5	64.1	60.9	99.3	93.2	70.3
w/o Grounding Loss	57.3	81.3	33.4	97.6	89.5	47.2
w/o SEG/REJ Weight	59.3	55.8	62.7	99.4	92.9	67.5
w/o R-Insturct	50.4	2.3	98.6	86.3	88.9	80.3
w/o R-Instruct-A	60.0	59.8	60.2	99.0	92.2	74.9
w/o R-Instruct-B	58.5	53.8	63.2	99.6	93.8	60.1

Table 6: Ablation studies on different modules.

LLaVA-13B, and ChatGPT-3.5. Moreover, VBackChecker achieves state-of-the-art performance among open-source models with a simpler end-to-end framework, approaching the performance of GPT-4o. VBackChecker is also evaluated on the POPE benchmark , which consists solely of simple object queries. Even in this relatively less challenging setting for hallucination detection, it consistently demonstrates superior performance, further highlighting its strong generalization capability across different hallucination detection tasks. These results showcase our VBackChecker as a highly effective hallucination checker, offering an efficient, and accurate solution for multimodal applications.

Ablation Studies

Table 6 validates our method's components. Isolating the grounding effect (Row 2) yields a 5.2% improvement over an instruction-tuned LLaVA baseline on R^2-HalBench. Emphasizing [SEG] and [REJ] token learning (Row 3) further aids in distinguishing grounded vs. hallucinated responses. Rows 4-6 confirm the value of our instruction data, particularly rich-context descriptions with and without masks.

To evaluate generalization, we analyze performance across source models and response lengths in Table 5. VBackChecker maintains consistent detection performance across diverse MLLM sources (Left). Furthermore, it remains reliable across varying lengths (1–110 words), including extensive rich-context responses (31–110 words) as shown in Table 5 (Right), validating its robustness for complex multimodal reasoning.

Conclusion

With the principle of "Seeing is Believing," we designed VBackChecker, a novel reference-free hallucination detection framework that handles rich-context responses and offers interpretability through grounding back, without relying on external references or experts. In support of this, we presented an innovative pipeline to generate instruction-tuning data (R-Instruct). We also established R^2-HalBench, a new benchmark reflecting real-world hallucination detection challenges. Experimental results demonstrated that VBackChecker achieves state-of-the-art performance on R^2-HalBench, rivaling GPT-4o's capabilities, while substantially improving pixel-level grounding accuracy. We expect these contributions to advance development of more reliable multimodal systems for real-world deployment.

Acknowledgements

This work was supported by National Natural Science Foundation of China (No.62576109, 62072112).

References

Achiam, J.; Adler, S.; Agarwal, S.; Ahmad, L.; Akkaya, I.; Aleman, F. L.; Almeida, D.; Altenschmidt, J.; Altman, S.; Anadkat, S.; et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35: 23716–23736.

Bai, Z.; Wang, P.; Xiao, T.; He, T.; Han, Z.; Zhang, Z.; and Shou, M. Z. 2024. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Chen, K.; Zhang, Z.; Zeng, W.; Zhang, R.; Zhu, F.; and Zhao, R. 2023. Shikra: Unleashing Multimodal LLM's Referential Dialogue Magic. *arXiv preprint arXiv:2306.15195*.

Chen, X.; Wang, C.; Xue, Y.; Zhang, N.; Yang, X.; Li, Q.; Shen, Y.; Liang, L.; Gu, J.; and Chen, H. 2024a. Unified hallucination detection for multimodal large language models. *arXiv preprint arXiv:2402.03190*.

Chen, Y.-C.; Li, W.-H.; Sun, C.; Wang, Y.-C. F.; and Chen, C.-S. 2024b. SAM4MLLM: Enhance Multi-Modal Large Language Model for Referring Expression Segmentation. In *European Conference on Computer Vision*, 323–340. Springer.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024c. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

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

Jing, L.; Li, R.; Chen, Y.; Jia, M.; and Du, X. 2023. FAITHSCORE: Evaluating Hallucinations in Large Vision-Language Models. *arXiv preprint arXiv:2311.01477*.

Kirillov, A.; Mintun, E.; Ravi, N.; Mao, H.; Rolland, C.; Gustafson, L.; Xiao, T.; Whitehead, S.; Berg, A. C.; Lo, W.-Y.; et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 4015–4026.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2023. Lisa: Reasoning segmentation via large language model. *arXiv preprint arXiv:2308.00692*.

Lai, X.; Tian, Z.; Chen, Y.; Li, Y.; Yuan, Y.; Liu, S.; and Jia, J. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9579–9589.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

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

Liu, C.; Ding, H.; and Jiang, X. 2023. Gres: Generalized referring expression segmentation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 23592–23601.

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

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023b. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023c. Visual instruction tuning. *Advances in neural information processing systems*, 36: 34892–34916.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023d. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*.

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

Panagopoulou, A.; Xue, L.; Yu, N.; Li, J.; Li, D.; Joty, S.; Xu, R.; Savarese, S.; Xiong, C.; and Niebles, J. C. 2024. X-InstructBLIP: A Framework for Aligning Image, 3D, Audio, Video to LLMs and its Emergent Cross-Modal Reasoning. In *European Conference on Computer Vision*, 177–197. Springer.

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

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Rasheed, H.; Maaz, M.; Shaji, S.; Shaker, A.; Khan, S.; Cholakkal, H.; Anwer, R. M.; Xing, E.; Yang, M.-H.; and Khan, F. S. 2024. Glamm: Pixel grounding large multimodal model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13009–13018.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Ren, T.; Liu, S.; Zeng, A.; Lin, J.; Li, K.; Cao, H.; Chen, J.; Huang, X.; Chen, Y.; Yan, F.; Zeng, Z.; Zhang, H.; Li, F.; Yang, J.; Li, H.; Jiang, Q.; and Zhang, L. 2024. Grounded SAM: Assembling Open-World Models for Diverse Visual Tasks. arXiv:2401.14159.

Wang, J.; Wang, Y.; Xu, G.; Zhang, J.; Gu, Y.; Jia, H.; Wang, J.; Xu, H.; Yan, M.; Zhang, J.; et al. 2023a. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Wang, J.; Wang, Y.; Xu, G.; Zhang, J.; Gu, Y.; Jia, H.; Yan, M.; Zhang, J.; and Sang, J. 2023b. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

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

Wang, P.; Bai, S.; Tan, S.; Wang, S.; Fan, Z.; Bai, J.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; et al. 2024a. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Wang, X.; Zhang, S.; Li, S.; Kallidromitis, K.; Li, K.; Kato, Y.; Kozuka, K.; and Darrell, T. 2024b. SegLLM: Multi-round Reasoning Segmentation. *arXiv preprint arXiv:2410.18923*.

Xia, Z.; Han, D.; Han, Y.; Pan, X.; Song, S.; and Huang, G. 2024. GSVA: Generalized Segmentation via Multimodal Large Language Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 3858–3869.

Yang, S.; Qu, T.; Lai, X.; Tian, Z.; Peng, B.; Liu, S.; and Jia, J. 2023. An improved baseline for reasoning segmentation with large language model. *arXiv e-prints*, arXiv–2312.

Zhang, S.; Sun, P.; Chen, S.; Xiao, M.; Shao, W.; Zhang, W.; Chen, K.; and Luo, P. 2023. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*.

Zhang, Y.; Huang, X.; Ma, J.; Li, Z.; Luo, Z.; Xie, Y.; Qin, Y.; Luo, T.; Li, Y.; Liu, S.; et al. 2024. Recognize anything: A strong image tagging model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 1724–1732.

Zhu, D.; Chen, J.; Shen, X.; Li, X.; and Elhoseiny, M. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*.