MHALO: Evaluating MLLMs as Fine-grained Hallucination Detectors

Anonymous ACL submission

Abstract

Hallucination remains a critical challenge for multimodal large language models (MLLMs), undermining their reliability in real-world applications. While fine-grained hallucination detection (FHD) holds promise for enhancing high-quality vision-language data construction and model alignment through enriched feedback signals, automated solutions for this task have yet to be systematically explored. Inspired by the concept of "MLLM as a Judge", we introduce MHALO, the first comprehensive benchmark specifically designed for evaluating MLLMs' capability in performing token-level FHD. Our benchmark encompasses 12 distinct hallucination types spanning both multimodal perception and reasoning domains. Through extensive evaluations of 9 selected MLLMs, we reveal substantial performance limitations, with the leading model achieving an average $\boldsymbol{F1_{IoU}}$ of only 40.59%. To address this limitation, we develop HALODET-4B, a specialized model trained on our curated training data, which significantly outperforms existing models. We hope the benchmark can provide valuable insights for future research on hallucination mitigation in MLLMs. The code and dataset will be publicly available.

1 Introduction

The advancement of Multimodal Large Language Models (MLLMs; (OpenAI, 2024; gpt, 2023; Team et al., 2024; Anthropic, 2024)) represents a groundbreaking achievement in the field of AI, demonstrating exceptional capabilities in perception and reasoning (Wang et al., 2024b; OpenAI, 2024; gpt, 2023; Team et al., 2024; Liu et al., 2024b). Despite their promise, MLLMs are still plagued by hallucination, a phenomenon that involves generating erroneous or fabricated responses contradicting the actual visual content or language context (Bai et al., 2024a; Liu et al., 2024a; Sahoo et al., 2024).

Therefore, to address this issue and enhance the reliability of MLLMs, **F**ine-grained **H**allucination **D**etection (FHD), which offers enriched token-level feedback signals, emerges as a crucial solution to mitigate the generation of erroneous or fabricated responses. Unlike coarse-grained feedback that penalizes hallucinations at the expense of suppressing correct content (Yu et al., 2024), FHD accelerates human annotation and data refinement by pinpointing hallucinations (Fu et al., 2024b), thereby facilitating efficient acquisition of high-quality data. Furthermore, FHD offers more informative signals, leading to effective model alignment. (Yu et al., 2024; Gunjal et al., 2024; Jing and Du, 2024a; Xiao et al., 2024).

Despite its advantages, current research on multimodal hallucination detection still exhibits limitations in granularity. Jing et al. (2024) was one of the first to conduct fine-grained hallucinations evaluations by verifying the extracted atomic facts in responses against the input image. Chen et al. (2024b) proposed a unified detection framework using external tools to validate hallucinations. Both of them operate detection at the claim level and lack the ability to precisely localize hallucinations. Automated hallucination detection at a more fine-grained level, token level, remains unexplored.

Inspired by the concept of MLLMs as a judge (Lee et al., 2024; Chen et al., 2024a; Wen et al., 2024), a natural question emerges: *"Can MLLMs serve as reliable judges for FHD?"* This necessitates establishing a meta-evaluation benchmark that can effectively assess the performance of MLLMs on FHD. Building such a benchmark presents two key challenges: (1) Construct a tailored dataset ensuring **comprehensive coverage of hallucination types** across diverse scenarios. (2) Developing **quantitative and objective evaluation metrics** that align with human judgment.

To bridge these research gaps, we introduce **MHALO**, a novel FHD benchmark consisting of 2,155 carefully curated instances with token-level annotations. It features the following aspects: on

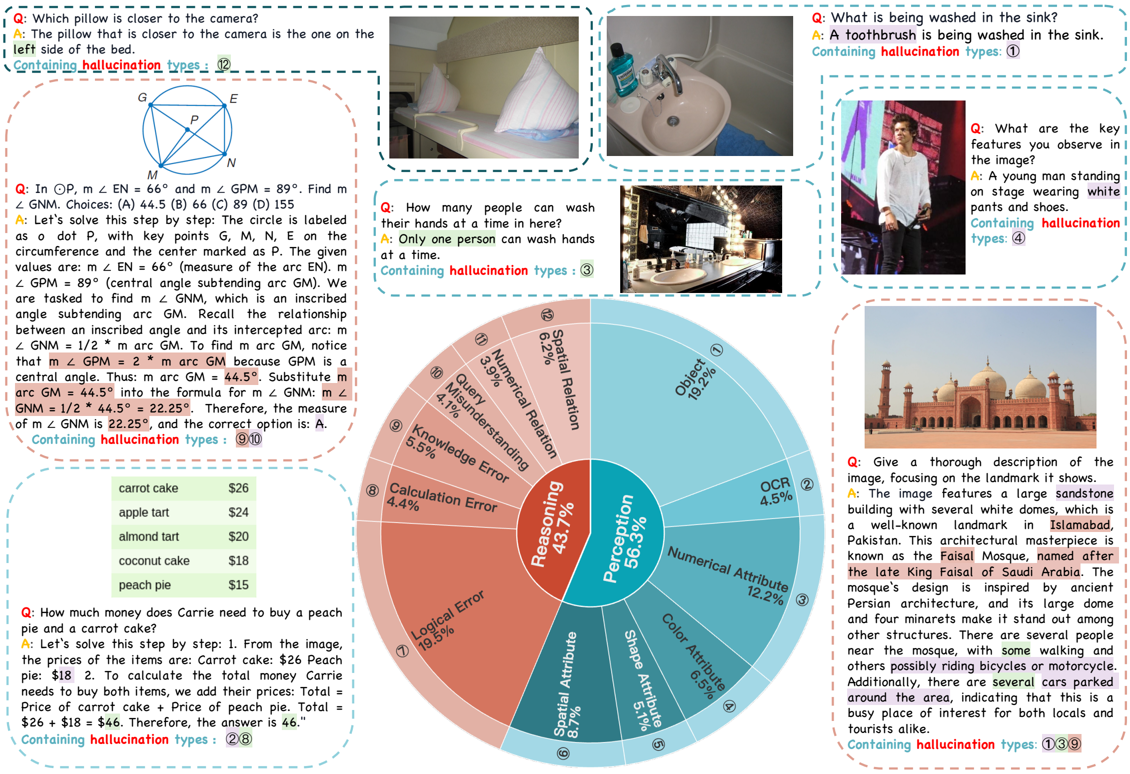


Figure 1: **The MHALO Benchmark.** Our benchmark features token-level annotations with comprehensive coverage of hallucination types across both Perception and Reasoning scenarios. Text highlighted in different colors corresponds to various types of hallucination annotations.

the one hand, prior work has mainly focused on natural scenes and contains only a small proportion of questions requiring mathematical reasoning (Chen et al., 2024b; Yu et al., 2024; Gunjal et al., 2024), leaving a comprehensive investigation of hallucination detection within vision-language reasoning largely unexplored. Thus, we present a comprehensive taxonomy covering hallucinations in both multimodal perception and reasoning processes, categorizing hallucinations into 12 distinct types (see the pie chart in Figure 1). On the other hand, MHALO defines FHD as a task requiring models to provide token-level hallucination annotations (see examples in Figure 1), taking into account both recognition and localization aspects, and we propose the corresponding metrics $\boldsymbol{F1_M}$ and $\boldsymbol{F1_{IoU}}$, the latter inspired by object detection to objectively assess the accuracy of detection. We demonstrate their effectiveness through rigorous validation.

We evaluate multiple well-known MLLMs (OpenAI, 2024; Anthropic, 2024) on MHALO and investigate the impact of different prompting strategies on their performance. It can be observed that FHD poses significant challenges for state-of-the-art (SOTA) MLLMs, with the leading MLLM on MHALO, GPT-4o, achieving an average $\boldsymbol{F1_{IoU}}$ of only 40.59%. In order to build a high-performance fine-grained hallucination detector, we adopt a data-driven strategy to fine-tune a specialized model HALODET-4B, which achieves SOTA performance on MHALO. Our contributions are as follows:

1. We propose a comprehensive FHD benchmark covering hallucination types both in perception and reasoning scenarios with specially-designed metrics $\boldsymbol{F1_M}$ and $\boldsymbol{F1_{IoU}}$ for token-level hallucination.

2. In our benchmark evaluation of various MLLMs, we have identified a significant performance gap in executing FHD. Notably, none of the models have surpassed the 50% threshold in terms of $\boldsymbol{F1_{IoU}}$.

3. We develop HALODET-4B, a detector that achieves SOTA performance on the proposed benchmark.

2 MHALO

We present MHALO, a novel benchmark encompassing 2,155 meticulously curated entries. The

Type	Definition
Object	Incorrect identification of objects in visual content.
OCR	Failure in text recognition processes within images.
Numerical Attribute	Misinterpretation of numerical values in visual elements.
Color Attribute	Errors in identifying the color.
Shape Attribute	Misrecognition of object shapes.
Spatial Attribute	Errors in recognizing the position, orientation, or distance of the object.
Logical Error	Errors in reasoning, such as incorrect causal relationships or conflicts in inference steps.
Calculation Error	Errors in mathematical operations (e.g., addition, subtraction, equation solving).
Knowledge Error	Applies incorrect domain knowledge or makes unrealistic inferences (e.g., violating common sense or physical laws).
Query Misunderstanding	Provides incorrect or irrelevant answers due to misunderstanding the query.
Numerical Relation	Misinterpreting the numerical relationship between objects (e.g., misreading proportions or quantities).
Spatial Relation	Misunderstanding the spatial, orientation, or distance relationships between objects.

Table 1: Hallucination types and definitions

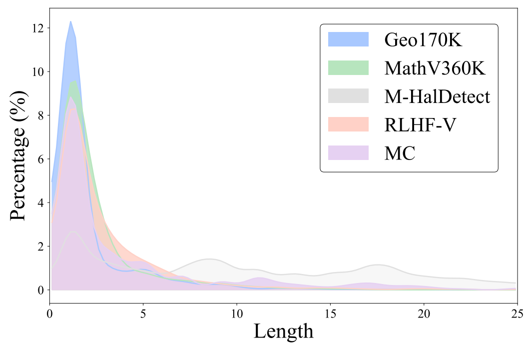


Figure 2: Fine-grained Annotation. Hallucinated segment length distribution of different subsets.

benchmark construction addresses three core challenges: (1) granular annotation framework, (2) comprehensive hallucination taxonomy, and (3) detection-oriented metrics. We begin by describing its unique features compared to previous works (Section 2.1). Next, we outline the data curation pipeline (Section 2.2). Finally, we show the design of token-level metrics (Section 2.3) for the automated and quantitative assessment of precise detection performance.

2.1 Key Features of MHALO

Our benchmark advances previous work through two fundamental innovations:

Token-level Annotation Framework. Unlike the existing detection approaches supporting claim-level (Chen et al., 2024b; Jing et al., 2024), which require extracting claims from annotated text, MHALO offers a token-level hallucination annotation directly on the predicted response, as illustrated in Figure 1. We ensure the annotation identifies the minimal erroneous components requiring revision during the dataset construction process, more details can be found in Appendix A. The distribution of hallucinatory segment lengths in our benchmark is shown in Figure 2. Most segments have fewer than five tokens, highlighting the precise localization of the hallucinatory part rather than offering a rough and approximate annotation. This enables accurate token-level feedback, facilitating the data selection process and enhancing the post-training process through fine-grained reward techniques (Yu et al., 2024; Gunjal et al., 2024).

Unified Perception-Reasoning Taxonomy. Reasoning is indispensable to fully unlocking the potential of Artificial General Intelligence (AGI) (Wang et al., 2023b). Earlier studies predominantly focused on hallucinations in natural scenes (Chen et al., 2024b; Yu et al., 2024; Gunjal et al., 2024), with only a limited proportion of questions involving mathematical reasoning. Nevertheless, actually, hallucinations can occur in both perception and reasoning processes. As shown in Figure 1 and Table 1, we distinguish two hierarchical stages:

- **Perception** involving image understanding and information extraction (e.g., misinterpretations of objects, text, or visual attributes like color, shape, and spatial positioning).
- **Reasoning** builds upon perception to infer relationships between objects or interpret complex scenarios (e.g., logical fallacies, computational errors, or misinterpretations of complex queries).

Our taxonomy identifies 12 distinct hallucination types across both stages beyond conventional hallucination types like object and attribute errors (Bai et al., 2024b; Jiang et al., 2024), enabling holistic evaluation of MLLMs in diverse scenarios.

Statistic	Number
NATURE	1000
RLHF-V	500
M-HalDetect	500
REASONING	1000
Geo170K	500
MathV360K	500
MC	155
Total	2155
Average hallucinated segment length	3.45
Average response length	72.41

Table 2: Detailed statistics of MHALO

2.2 Dataset Collection Process

MHALO comprises instance tuples (I, Q, O, A), where I denotes the input image, Q represents the query prompt, O indicates the potentially hallucinated response, and A serves as the ground truth annotation with fine-grained hallucination tags for each hallucinated segments. The dataset of MHALO can be divided into three distinct subsets: (1) The **NATURE** set is curated from two existing human-labeled fine-grained hallucination datasets (Yu et al., 2024; Gunjal et al., 2024), and we further filter and tailor it to meet requirements of the benchmark. (2) The **REASONING** set focus on reasoning. As most existing multimodal mathematical reasoning datasets are coarsely annotated, we adopt the way of perturbing ground-truth solutions (Fu et al., 2024a; Mishra et al., 2024b) to acquire large amounts of fine-grained annotated instances. (3) To further verify the detector's performance in real-world applications, we collect out-of-distribution datasets (Sun et al., 2023; Lu et al., 2024; Guan et al., 2024) covering both perception and reasoning aspects and apply manual fine-grained annotation, denoted as the **MC** set.

Accordingly, the NATURE set evaluates the hallucination detection ability mainly from perception aspects, the REASONING set stresses whether MLLM evaluators can truly detect hallucination in a multimodal reasoning process. The detailed statistics are shown in Table 2. We provide the detailed construction process of each subset in Appendix A.

Quality Examination. To ensure the accuracy and granularity of hallucination annotations, we manually evaluated the dataset. Three authors independently reviewed a 200-entry sample from the benchmark. The success rate was determined by majority voting, considering a sample successful only if at least two annotators agreed on its fine-grained annotation quality. The evaluation results revealed a success rate of 95%, supported by a substantial inter-annotator agreement of 0.79, as measured by Fleiss' Kappa (Fleiss et al., 1981). These findings validate the high quality of our dataset. Further details can be found in Appendix A.4.

2.3 Metric

FHD Task: Given a multimodal query q consisting of an image I and a textual prompt Q, and the corresponding output O from an MLLM, our task is to identify and localize all hallucinated intervals in O, as shown in Figure 3. Hallucinated intervals are text segments in O that are not grounded in the input query q.

Notation and Definitions: To formalize this task, we introduce the following notations:

- $\mathcal{G} = \{B^1_{\text{gt}}, B^2_{\text{gt}}, \ldots, B^m_{\text{gt}}\}$: Ground truth intervals, where $B^j_{\text{gt}} = [g_j, h_j]$ for $j = 1, 2, \ldots, m$. Here, g_j and h_j represent the start and end token indices of the j-th ground truth interval in the sequence of tokens $O = [o_1, o_2, \ldots, o_n]$.

- $\mathcal{O} = \{B^1_p, B^2_p, \ldots, B^n_p\}$: Predicted intervals, where $B^i_p = [s_i, t_i]$ for $i = 1, 2, \ldots, n$. The indices s_i and t_i denote the start and end token indices of the i-th predicted interval.

- $T(B)$: The text span corresponds to an interval B in O, where B can refer to either a ground truth or a predicted interval. This is the actual sequence of tokens within the indices defined by the interval.

We use two metrics to evaluate the model's performance: $\boldsymbol{F1_M}$ and $\boldsymbol{F1_{IoU}}$.

▷ $\boldsymbol{F1_M}$: The evaluation of the model's performance is based on partial matches between the ground truth intervals and the predicted intervals. Specifically, we use a recall-based partial match score (PM_R) (Jafari et al., 2024) to assess the degree to which the predicted intervals match the ground truth intervals. PM_R is defined as:

$$PM_R(j) = \begin{cases} 1, & \text{if } \exists B^i_p \text{ s.t. } B^i_p = B^j_{\text{gt}}, \\ \frac{|T(B^i_p)|}{|T(B^j_{\text{gt}})|}, & \text{if } \exists B^i_p \text{ s.t. } B^i_p \subseteq B^j_{\text{gt}}, \\ 0, & \text{otherwise}. \end{cases} \tag{1}$$

Similarly, the precision-based partial match score PM_P is defined analogously. The recall Rec_M

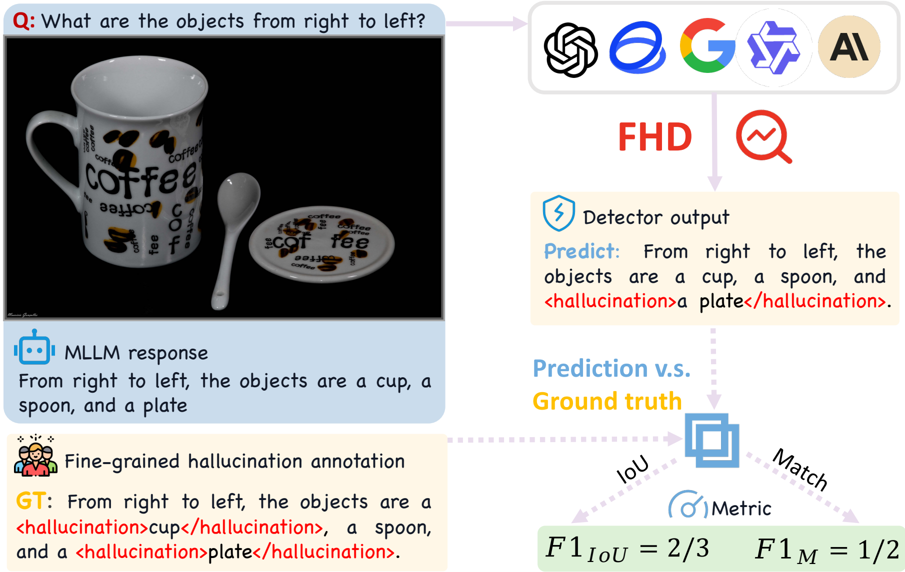


Figure 3: An overview of token-level FHD and corresponding metrics.

and precision $Prec_M$ are then computed as:

$$Rec_M = \frac{1}{m} \sum_{j=1}^{m} PM_R(j), \quad (2)$$

$$Prec_M = \frac{1}{n} \sum_{i=1}^{n} PM_P(i). \quad (3)$$

Finally, $\boldsymbol{F1_M}$ is calculated as the harmonic mean of recall and precision.

▷ $\boldsymbol{F1_{IoU}}$: Although $\boldsymbol{F1_M}$ can indicate the degree of overlap between predictions and ground truth, it fails to capture the **inherent ambiguity** in detection tasks. For example, when annotating hallucinations, there may be multiple valid ways to label the text. For instance, an image showing a black shirt and white pants could lead to MLLM hallucination responses like "black pants". Both "black" and "pants" could be valid hallucinations, where simply measuring the proportion of matched tokens becomes less meaningful. Inspired by object detection metrics (Padilla et al., 2020; Zang et al., 2022), we propose $\boldsymbol{F1_{IoU}}$ to mitigate this issue. First, we introduce the Intersection over Union (IoU) score, which measures the overlap between predicted and ground truth intervals. The IoU is defined as:

$$\text{IoU}(i, j) = \frac{|B_p^i \cap B_{\text{gt}}^j|}{|B_p^i \cup B_{\text{gt}}^j|}. \quad (4)$$

A match is considered valid if IoU ≥ 0.5. Let $\mathbb{1}(\cdot)$ be the indicator function, the $\boldsymbol{F1_{IoU}}$ score is computed through optimal interval matching:

$$\widehat{M} = \max_{\mathcal{M} \in \mathbb{M}} \sum_{(i,j) \in \mathcal{M}} \mathbb{1}(\text{IoU}(i, j) \geq 0.5) \quad (5)$$

$$\boldsymbol{F1_{IoU}} = \frac{2\widehat{M}}{|\mathcal{O}| + |\mathcal{G}|} \quad (6)$$

where $\mathbb{M} := \{\mathcal{M} \subseteq \mathcal{O} \times \mathcal{G} \mid \forall (a, b), (c, d) \in \mathcal{M}, \ (a \neq c) \wedge (b \neq d)\}$ is the set of all bipartite matchings, and $\widehat{M}$ is the maximum matching solved by the Hungarian algorithm (Kuhn, 1955). In this way, we anticipate a more precise evaluation of the detection of hallucinatory segments.

3 Fine-tuning an MLLM as a Detector

In our preliminary experiments, we observed that leading MLLMs (OpenAI, 2024; Team et al., 2024) are not particularly effective at detecting hallucinatory segments (see Table 3). This shortfall is probably due to the absence of such tasks in the training data, which has prevented the full potential of these models from being realized. To enhance the ability of hallucination detection, we collect and construct labeled data and train a specialized detection model. Specifically, we use GLM-4V (4B) (GLM et al., 2024) as our backbone model and fine-tune it to get HaloDet-4B. The training set is constructed using a process similar to Section 2.2. Additional details about training set construction and fine-tuning parameters can be found in Appendix B.1.

MLLM	RLHF-V			M-HalDetect			Geo170K			MathV360K			MC			Average		
	$F1_M$	$F1_{IoU}$	IF	$F1_M$	$F1_{IoU}$	IF	$F1_M$	$F1_{IoU}$	IF	$F1_M$	$F1_{IoU}$	IF	$F1_M$	$F1_{IoU}$	IF	$F1_M$	$F1_{IoU}$	IF
Open-Source Evaluation Models																		
MINICPM-V 2.6	25.09	20.96	99.20	8.53	3.01	97.00	10.73	3.90	41.83	24.48	19.15	88.74	34.97	32.42	83.33	18.36	13.13	82.14
INTERNVL2-LLAMA3-76B	31.62	26.81	99.59	15.00	6.30	99.53	28.90	19.44	58.46	32.49	25.53	94.38	50.53	44.57	98.47	28.54	21.17	88.09
LLAMA-3.2-90B-VISION-INSTRUCT	35.71	29.49	99.79	18.40	7.81	99.80	36.85	18.36	79.76	42.54	32.72	96.78	55.49	45.52	95.30	34.89	23.63	94.05
Closed-Source Evaluation Models																		
QWEN-VL-MAX	32.70	26.66	100.00	11.43	9.65	100.00	37.55	19.65	99.39	37.02	31.95	99.60	40.19	34.98	98.67	30.38	22.89	99.67
ABAB7-CHAT-PREVIEW	38.38	32.42	98.99	27.94	16.39	96.99	34.16	19.33	88.41	40.57	35.79	95.40	57.64	53.33	98.70	36.88	27.97	95.23
GLM-4V-PLUS	38.85	32.30	99.80	28.65	20.38	99.80	34.91	30.83	81.40	37.33	33.99	94.78	48.12	42.36	97.40	35.87	30.30	94.19
CLAUDE-3.5-SONNET	43.94	28.73	99.00	39.65	20.59	97.80	51.69	27.36	98.80	56.87	37.77	98.20	58.32	44.91	99.29	48.71	29.68	98.50
CLAUDE-3.5-SONNET*	43.98	28.12	99.80	44.46	25.24	98.60	52.02	26.76	98.40	56.08	35.60	99.00	59.02	47.16	97.87	49.79	30.13	98.88
GEMINI-1.5-PRO	41.54	29.71	99.60	35.83	19.64	99.80	52.96	30.17	99.60	62.01	50.79	100.00	63.90	55.50	99.35	49.22	34.22	99.72
GEMINI-1.5-PRO*	44.37	33.00	99.60	37.14	21.09	99.60	56.00	31.07	98.40	65.27	54.29	98.80	68.03	60.07	98.05	51.94	36.67	99.03
GPT-4O	43.92	30.63	100.00	45.97	32.85	100.00	63.03	45.22	98.80	62.63	45.12	99.80	64.90	56.07	99.29	54.63	39.62	99.63
GPT-4O*	46.55	35.74	99.80	47.27	30.30	100.00	57.77	34.08	98.80	72.61	58.27	99.00	**70.97**	58.69	96.52	56.83	40.59	99.24
HALODET-4B	**49.11**	**39.70**	99.00	**56.49**	**47.06**	99.80	**70.64**	**61.43**	96.20	**73.39**	**64.54**	95.00	70.77	**61.31**	98.70	**63.01**	**53.76**	97.59

Table 3: The overall performance of different MLLMs on MHALO (%). The best results are highlighted in **bold**, while the suboptimal ones are marked with underline. Models using Analyze-then-Judge prompting are denoted with *.

4 Experiments

4.1 Experimental Setup

Model Selection. We evaluate a total of 10 MLLMs on MHALO, including GPT-4O (OpenAI, 2024), GEMINI-1.5-PRO-002 (Team et al., 2024), CLAUDE-3-5-SONNET (Anthropic, 2024), GLM-4V-PLUS (GLM et al., 2024), ABAB7-CHAT-PREVIEW[1], QWEN-VL-MAX (Bai et al., 2023), LLAMA-3.2-90B-VISION (AI@Meta, 2024), INTERNVL2-LLAMA3-76B (Chen et al., 2024c), MINICPM-V-2.6 (Yao et al., 2024), and our trained expert detector HALODET-4B.

Evaluation Metrics. We utilize the metrics $F1_{IoU}$ and $F1_M$ defined in section 2.3. Given the challenges faced by MLLM in performing FHD, the testee models sometimes fail to follow the instruction. We introduce the metric IF to represent the proportion of successful entries that complete the FHD task, samples on which the model fails to accomplish the task will receive a score of zero for these metrics.

Evaluation Settings. We experiment various prompting strategies to evaluate the testee models: (1) The baseline method uses direct instructions to prompt the MLLM for FHD in a zero-shot setting. The MLLM then outputs the detection result using XML-style tags, as illustrated in Figure 3. We provide the discussion of using different annotation formats in Appendix C.2. (2) To further explore the capability of MLLMs to perform FHD, we experiment with three additional prompting strategies (See Appendix C.1 for details). Our results indicate that the "Analyze-then-Judge" paradigm achieves superior performance across nearly all subsets. It builds on prior one-step chain-of-thought evaluation (Chiang and yi Lee, 2023; Wei et al., 2023; Chen et al., 2024a), and we implement it through a two-phase reasoning process that first generates a detailed hallucination analysis with factual corrections and then annotating the response with hallucination tags. Here, we evaluate all the models using the baseline method and also evaluate the performance of SOTA MLLMs with "Analyze-then-Judge". The prompts used for evaluation are provided in Table 11 and Table 14 in Appendix D.

4.2 Main Results

The results of ten selected MLLMs on MHALO are presented in Table 3. Our comprehensive evaluation yields the following key insights:

FHD remains a challenge for SOTA MLLMs. Despite significant advancements in current MLLMs, top-performance models still struggle with FHD. The results show that GPT-4O leads the benchmark, but achieves an average $F1_{IoU}$ of only 40.59%, and GEMINI-1.5-PRO follows behind. Notably, nearly half of the evaluated models exhibit particularly weak performance, with $F1_{IoU}$ values below 30%, especially among open-source models, which exhibit the worst results. These findings highlight inherent limitations in their capabilities for FHD.

Lightweight HALODET-4B achieves superior performance. HALODET-4B outperforms the best commercial model, GPT-4O, by an impressive margin, achieving nearly a 13% absolute gain in average $F1_{IoU}$. Furthermore, it nearly achieves SOTA performance across all the subsets. These results underscore the critical need for specialized solutions like HALODET-4B, while also highlighting the substantial room for improvement in general-

[1] https://www.minimaxi.com/en/news/abab7-preview-release.

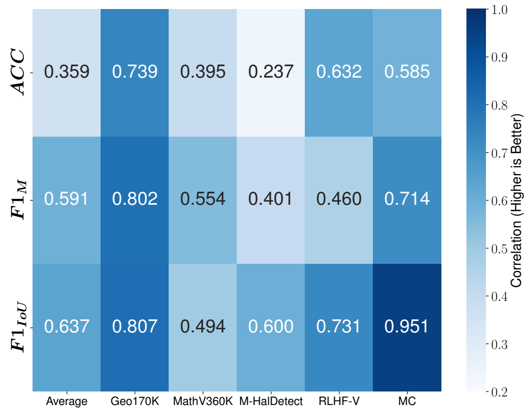


Figure 4: Metric Correlation with Human Evaluation.

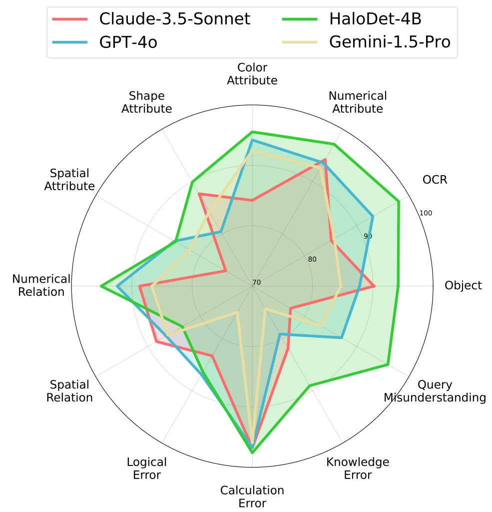


Figure 5: Detection Performance of four representative MLLMs across 12 hallucination types.

purpose MLLMs for FHD. A detailed case study comparing the detection outputs of various models can be found in Appendix E.

5 Analysis

5.1 Metric Correlation with Human Evaluation

To evaluate whether $\boldsymbol{F1_{IoU}}$ and $\boldsymbol{F1_M}$ can serve as reliable proxies for human judgment, we compute their Pearson correlation coefficients (Cohen et al., 2009) with human annotations across MHALO. Three authors independently scored each predicted hallucination segment in GPT-4o detection results using an integer scale x $(1 \leq x \leq 4)$, which reflects the degree of correctness and precision in identifying hallucinated segments. The details criteria are provided in Table 4 in Appendix B.2. The final score for each sample is obtained by averaging the scores of all predicted hallucination segments. We compare our metrics against token-level accuracy (*ACC*) from (Fu et al., 2024b), which formulates hallucination detection as a binary token classification task.

The overall results are presented in Figure 4. $\boldsymbol{F1_{IoU}}$ demonstrates the strongest alignment with human judgments, achieving Pearson correlation scores of 0.951 and 0.807 on the MC and Geo170K datasets, respectively. In contrast, $\boldsymbol{F1_M}$ exhibits suboptimal alignment, while *ACC* shows significantly weaker correlations, with an overall correlation score of just 0.359. We attribute this discrepancy to the following factors: (1) $\boldsymbol{F1_{IoU}}$ and $\boldsymbol{F1_M}$ explicitly account for the spatial alignment of intervals, while *ACC* reduces detection to binary token classification, which fails to capture the granularity of the annotations. (2) $\boldsymbol{F1_{IoU}}$ utilizes thresholding for interval matching, effectively addressing annotation boundary ambiguity and enhancing both flexibility and robustness across diverse datasets, thus achieving superior performance compared to $\boldsymbol{F1_M}$. $\boldsymbol{F1_{IoU}}$ proves to be the most reliable proxy for human judgment, followed by $\boldsymbol{F1_M}$, whereas traditional token-level metrics, such as *ACC*, exhibit significant limitations in the FHD task.

5.2 Performance in Identifying Different Types of Hallucinations

Figure 5 presents the accuracies of cutting-edge models on MHALO in identifying hallucinations across different types. The models evaluated include GPT-4O (OpenAI, 2024), CLAUDE-3-5-SONNET (Anthropic, 2024), GEMINI-1.5-PRO (Team et al., 2024), and HALODET-4B. We provide the experiment details in Appendix B.3. We can observe that MLLMs excel at identifying hallucinations involving **numerical attribute and calculation error**, achieving over 90% accuracy. However, they exhibit notable weaknesses with **logical error and spatial attribute**, which require advanced reasoning and spatial comprehension. While our HALODET-4B achieves a more balanced performance overall, it still struggles with spatial attribute and spatial relation. In summary, MLLMs are doing well in hallucinations related to arithmetic and object recognition, but face persistent challenges in logical coherence, spatial reasoning, and complex attribute understanding.

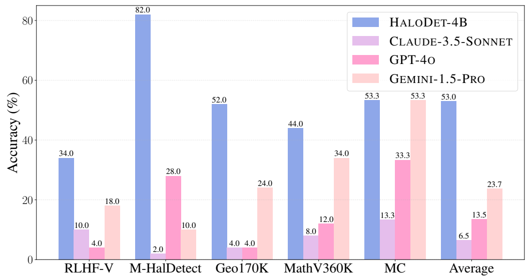


Figure 6: Results of four representative MLLMs on samples without hallucination in MHALO.

5.3 FHD on Non-Hallucinated Samples

Figure 6 shows the accuracy of various models in detecting hallucinations in non-hallucinated samples from MHALO. It reveals that current SOTA MLLMs tend to produce a high rate of false positives, incorrectly flagging truthful information as hallucinated, with average accuracy consistently below 25%. In contrast, our HaloDet-4B outperforms other MLLMs across all subsets. The performance gap is especially pronounced in the M-HalDetect dataset, where our method reaches an impressive 82%. On average, HaloDet-4B reaches 53% accuracy, more than twice the performance of the best-performing MLLM, highlighting its reliability.

6 Related Work

6.1 Hallucinations in MLLMs

Recent advances in MLLMs have achieved remarkable breakthroughs in cross-modal perception and reasoning (Wang et al., 2024b; OpenAI, 2024; gpt, 2023; Team et al., 2024; Liu et al., 2024b), enabling them to perform complex tasks requiring visual-language reasoning beyond basic recognition capabilities (GLM et al., 2024). Despite these advancements, hallucinations remain significant challenges, where MLLMs generate responses contradicting the visual input or linguistic context. This critical limitation hinders practical deployment like autonomous driving (Cui et al., 2024), where accurate and trustworthy performance is essential. Addressing this fundamental challenge is crucial to unlocking the full potential of MLLMs in real-world applications.

Previous studies have progressively expanded from initial investigations into object hallucinations (Rohrbach et al., 2019; Li et al., 2023) to the evaluation of a broader range of types involving category, attribute, and relation hallucinations (Bai et al., 2024a; Wang et al., 2024a; Jing et al., 2024). However, current research remains limited to natural scenarios, overlooking the critical dimensions of hallucinations induced during reasoning processes. In this paper, we bridge this gap by establishing a unified taxonomy that encompasses hallucination types across both the perception and reasoning stages.

6.2 MLLM as a Judge for Fine-grained Hallucination Detection

Recent advancements in hallucination evaluation and detection have moved towards a more fine-grained level, targeting evaluation at the sentence (Yan et al., 2024; Xiao et al., 2024), claim (Jing et al., 2024; Chen et al., 2024b), and even token levels (Jing and Du, 2024b). While the meta-evaluation paradigm, such as MLLM as a judge (Gu et al., 2024; Chen et al., 2024a; Lee et al., 2024), has yet to be systematically explored. For instance, Wang et al. (2023a) first proposed training MLLMs with synthetic data for hallucination detection, but their approach was limited to response level. Chen et al. (2024b) introduced a claim-level benchmark and suggested leveraging external tools to assist in hallucination detection. Nevertheless, this method is restricted to certain types of hallucinations, such as those involving factual knowledge or verifiable objects, leaving it ineffective in scenarios that require complex reasoning, such as identifying spatial relations. Additionally, claim-level detection requires extracting claims, which introduces further complexity. In this paper, we focus on exploring the potential of MLLMs to perform FHD at the token level.

7 Conclusion

In this paper, we introduce a novel meta-evaluation benchmark, MHALO, designed to assess different MLLMs ' capability in performing FHD. By systematically evaluating 9 well-known MLLMs, we highlight the significant performance gaps, none of the models exceeded 50% $\boldsymbol{F1_{IoU}}$. To address this limitation, we develop HaloDet-4B, a specialized model that significantly outperforms existing models. This benchmark, along with the trained detector, provides valuable tools for improving hallucination detection in MLLMs and can guide future research in model alignment.

8 Limitations

Our study makes progress in fine-grained hallucination detection for MLLMs through MHALO and HALODET-4B, but several limitations should be acknowledged to guide future research:

Optimization Potential of the Detector. Although HALODET-4B achieves SOTA performance on MHALO, our training employs standard hyperparameters without exhaustive optimization. The 4B parameter architecture is lightweight and efficient, but may not fully exploit the training data's potential. Systematic exploration of model scaling (e.g., 13B/70B variants), advanced optimization techniques, and architectural innovations could further boost detection accuracy.

Generalization Across Modalities. While MHALO covers 12 hallucination types, its current instantiation focuses on image-text interactions. Emerging multimodal scenarios involving video, audio, and 3D data may introduce new hallucination patterns requiring framework adaptation. Extending our methodology to these domains remains an open challenge.

References

2023. Gpt-4v(ision) system card.

AI@Meta. 2024. Llama 3 model card.

AI Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku. *Claude-3 Model Card*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024a. Hallucination of multimodal large language models: A survey. *Preprint*, arXiv:2404.18930.

Zechen Bai, Pichao Wang, Tianjun Xiao, Tong He, Zongbo Han, Zheng Zhang, and Mike Zheng Shou. 2024b. Hallucination of multimodal large language models: A survey. *arXiv preprint arXiv:2404.18930*.

Dongping Chen, Ruoxi Chen, Shilin Zhang, Yinuo Liu, Yaochen Wang, Huichi Zhou, Qihui Zhang, Yao Wan, Pan Zhou, and Lichao Sun. 2024a. Mllm-as-a-judge: Assessing multimodal llm-as-a-judge with vision-language benchmark. *Preprint*, arXiv:2402.04788.

Xiang Chen, Chenxi Wang, Yida Xue, Ningyu Zhang, Xiaoyan Yang, Qiang Li, Yue Shen, Lei Liang, Jinjie Gu, and Huajun Chen. 2024b. Unified hallucination detection for multimodal large language models. *Preprint*, arXiv:2402.03190.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, 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*, pages 24185–24198.

Cheng-Han Chiang and Hung yi Lee. 2023. A closer look into automatic evaluation using large language models. *Preprint*, arXiv:2310.05657.

Israel Cohen, Yiteng Huang, Jingdong Chen, Jacob Benesty, Jacob Benesty, Jingdong Chen, Yiteng Huang, and Israel Cohen. 2009. Pearson correlation coefficient. *Noise reduction in speech processing*, pages 1–4.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, 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*, pages 958–979.

Qingxiu Dong, Lei Li, Damai Dai, Ce Zheng, Jingyuan Ma, Rui Li, Heming Xia, Jingjing Xu, Zhiyong Wu, Tianyu Liu, et al. 2022. A survey on in-context learning. *arXiv preprint arXiv:2301.00234*.

Joseph L Fleiss, Bruce Levin, Myunghee Cho Paik, et al. 1981. The measurement of interrater agreement. *Statistical methods for rates and proportions*, 2(212-236):22–23.

Deqing Fu, Ameya Godbole, and Robin Jia. 2024a. Scene: Self-labeled counterfactuals for extrapolating to negative examples. *Preprint*, arXiv:2305.07984.

Deqing Fu, Tong Xiao, Rui Wang, Wang Zhu, Pengchuan Zhang, Guan Pang, Robin Jia, and Lawrence Chen. 2024b. Tldr: Token-level detective reward model for large vision language models. *Preprint*, arXiv:2410.04734.

Jiahui Gao, Renjie Pi, Jipeng Zhang, Jiacheng Ye, Wanjun Zhong, Yufei Wang, Lanqing Hong, Jianhua Han, Hang Xu, Zhenguo Li, and Lingpeng Kong. 2023. G-llava: Solving geometric problem with multi-modal large language model. *Preprint*, arXiv:2312.11370.

Team GLM, Aohan Zeng, Bin Xu, Bowen Wang, Chenhui Zhang, Da Yin, Dan Zhang, Diego Rojas, Guanyu Feng, Hanlin Zhao, et al. 2024. Chatglm: A family of large language models from glm-130b to glm-4 all tools. *arXiv preprint arXiv:2406.12793*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, et al. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and

Tianyi Zhou. 2024. Hallusionbench: An advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. *Preprint*, arXiv:2310.14566.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. *Preprint*, arXiv:2308.06394.

Nazanin Jafari, James Allan, and Sheikh Muhammad Sarwar. 2024. Target span detection for implicit harmful content. *Preprint*, arXiv:2403.19836.

Chaoya Jiang, Hongrui Jia, Wei Ye, Mengfan Dong, Haiyang Xu, Ming Yan, Ji Zhang, and Shikun Zhang. 2024. Hal-eval: A universal and fine-grained hallucination evaluation framework for large vision language models. *Preprint*, arXiv:2402.15721.

Liqiang Jing and Xinya Du. 2024a. Fgaif: Aligning large vision-language models with fine-grained ai feedback. *Preprint*, arXiv:2404.05046.

Liqiang Jing and Xinya Du. 2024b. Fgaif: Aligning large vision-language models with fine-grained ai feedback. *arXiv preprint arXiv:2404.05046*.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. *Preprint*, arXiv:2311.01477.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval research logistics quarterly*, 2(1-2):83–97.

Seongyun Lee, Seungone Kim, Sue Hyun Park, Geewook Kim, and Minjoon Seo. 2024. Prometheus-vision: Vision-language model as a judge for fine-grained evaluation. *Preprint*, arXiv:2401.06591.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. *Preprint*, arXiv:2305.10355.

Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, and Piotr Dollár. 2015. Microsoft coco: Common objects in context. *Preprint*, arXiv:1405.0312.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024b. Mmbench: Is your multi-modal model an all-around player? *Preprint*, arXiv:2307.06281.

Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. 2024. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. In *International Conference on Learning Representations (ICLR)*.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024a. Fine-grained hallucination detection and editing for language models. *Preprint*, arXiv:2401.06855.

Abhika Mishra, Akari Asai, Vidhisha Balachandran, Yizhong Wang, Graham Neubig, Yulia Tsvetkov, and Hannaneh Hajishirzi. 2024b. Fine-grained hallucinations detections. *arXiv preprint*.

OpenAI. 2024. Hello gpt-4o.

Rafael Padilla, Sergio L Netto, and Eduardo AB Da Silva. 2020. A survey on performance metrics for object-detection algorithms. In *2020 international conference on systems, signals and image processing (IWSSIP)*, pages 237–242. IEEE.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2019. Object hallucination in image captioning. *Preprint*, arXiv:1809.02156.

Pranab Sahoo, Prabhash Meharia, Akash Ghosh, Sriparna Saha, Vinija Jain, and Aman Chadha. 2024. A comprehensive survey of hallucination in large language, image, video and audio foundation models. *Preprint*, arXiv:2405.09589.

Wenhao Shi, Zhiqiang Hu, Yi Bin, Junhua Liu, Yang Yang, See-Kiong Ng, Lidong Bing, and Roy Ka-Wei Lee. 2024. Math-llava: Bootstrapping mathematical reasoning for multimodal large language models. *Preprint*, arXiv:2406.17294.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liang-Yan Gui, Yu-Xiong Wang, Yiming Yang, Kurt Keutzer, and Trevor Darrell. 2023. Aligning large multimodal models with factually augmented rlhf. *Preprint*, arXiv:2309.14525.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Jiaqi Wang, Haiyang Xu, Ming Yan, Ji Zhang, and Jitao Sang. 2024a. Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *Preprint*, arXiv:2311.07397.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023a. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023b. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Yiqi Wang, Wentao Chen, Xiaotian Han, Xudong Lin, Haiteng Zhao, Yongfei Liu, Bohan Zhai, Jianbo Yuan, Quanzeng You, and Hongxia Yang. 2024b. Exploring the reasoning abilities of multimodal large language models (mllms): A comprehensive survey on emerging trends in multimodal reasoning. *arXiv preprint arXiv:2401.06805*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Xueru Wen, Xinyu Lu, Xinyan Guan, Yaojie Lu, Hongyu Lin, Ben He, Xianpei Han, and Le Sun. 2024. On-policy fine-grained knowledge feedback for hallucination mitigation. *arXiv preprint arXiv:2406.12221*.

Wenyi Xiao, Ziwei Huang, Leilei Gan, Wanggui He, Haoyuan Li, Zhelun Yu, Fangxun Shu, Hao Jiang, and Linchao Zhu. 2024. Detecting and mitigating hallucination in large vision language models via fine-grained ai feedback. *arXiv preprint arXiv:2404.14233*.

Siming Yan, Min Bai, Weifeng Chen, Xiong Zhou, Qixing Huang, and Li Erran Li. 2024. Vigor: Improving visual grounding of large vision language models with fine-grained reward modeling. *Preprint*, arXiv:2402.06118.

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. 2024. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, and Tat-Seng Chua. 2024. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *Preprint*, arXiv:2312.00849.

Yuhang Zang, Wei Li, Kaiyang Zhou, Chen Huang, and Chen Change Loy. 2022. *Open-Vocabulary DETR with Conditional Matching*, page 106–122. Springer Nature Switzerland.

A Details on Benchmark Construction

A.1 The NATURE Set

The NATURE set focuses on tasks related to the perception and comprehension of natural images. The data used in this set are derived from RLHF-V (Yu et al., 2024) and M-HalDetect (Gunjal et al., 2024), two existing fine-grained annotated hallucination datasets labeled by human. RLHF-V is a fine-grained human preference dataset, containing 5.7k QA and captioning samples, the image-instruction pairs are collected from diverse datasets, mostly from COCO, and two corresponding outputs (O_w, O_l) in each instance, hallucinated outputs O_l generated by diverse MLLM, such as InstructBLIP, Qwen, and LLaVA, the corresponding refined response O_w is written by people through fixing the hallucination span in O_l. We adopt the image-instruction pair and O_l as I, Q, O, and acquire A by comparing O_w and O_l. We then further manually check and filter to make sure O is correctly annotated in each instance. We split part of the dataset to be used in the benchmark while remaining to construct a training set. M-HalDetect focuses exclusively on captioning tasks and includes 12k training samples and 3k testing samples. Its images are collected from COCO-val2014 (Lin et al., 2015), and corresponding caption output is sampled from MLLMs and the hallucination segment is labeled with the "Inaccurate" class. We sample instances from the testing set and further adjust their format to match our benchmark.

A.2 The REASONING Set

The REASONING set expands the benchmark's scope beyond the natural scenario to include mathematical reasoning. We meticulously select two multimodal math reasoning datasets Geo170K (Gao et al., 2023) and MathV360K (Shi et al., 2024) as the data source. Unlike the NATURE set, they only contain ground truth solution in each instance and no hallucination responses exist. Inspired by Mishra et al. (2024a), We apply the perturbation method to get the hallucinated solution and corresponding annotation. Geo170K is a multimodal geometry dataset containing more than 170K geometric problem instances, and the answers to each problem have a detailed reasoning process. To introduce hallucination in the solution while balancing the distribution of different types of hallucination in our taxonomy, we prompt GPT-4o (OpenAI, 2024), which takes the image-instruction pair and original solution as input and is instructed to generate hallucinated solution accompanying annotation for 12 different types. MathV360K is a multimodal mathematical reasoning dataset containing 360K question-answer pairs from different domains thus covering diverse tasks requiring reasoning. However, it only has a final answer and lacks the intermediate reasoning step. So we first prompt GPT-4o to generate the Chain-of-thought (CoT) (Wei et al., 2023) solutions. Then apply a similar process like Geo170K to insert hallucination. The corresponding prompt template is in Appendix D. We finally filtered samples to ensure the balanced coverage of different hallucination types.

A.3 The MC Set

We construct a carefully human-annotated dataset comprising 155 samples, with 81 entries sourced from MMHAL-BENCH (Sun et al., 2023),58 entries from MathVista (Lu et al., 2024), and 16 entries from HallusionBench (Guan et al., 2024) to ensure comprehensive coverage of perceptual and reasoning capabilities. Both source datasets provide sample-level annotations indicating response correctness from various MLLMs. We specifically select responses flagged as erroneous for fine-grained annotation, focusing on two key criteria: (1) *Correctness*. The annotated text segment should contain hallucinatory content. (2) *Granularity*. The proportion of hallucinatory content within the annotated segment. To ensure the quality of the data, all the samples were manually annotated by the authors of this paper and subsequently refined through a comprehensive review process. We employed a two-phase approach to maintain consistency in annotation. In the first phase, each sample was independently annotated by three annotators, with the criteria of identifying the smallest erroneous components requiring revision. This method resulted in a relatively high inter-annotation agreement rate of 86%, where consistency was defined as an exact match of each labeled hallucination segment for each sample. Specifically, of the 155 newly collected question-answer pairs, only 21 entries showed discrepancies in the annotations. Then we employed a majority voting system, where multiple authors collaboratively decided whether to retain or adjust contentious annotations. This was achieved through team discussions, ensuring consensus was reached on each sample.

Score	Description
1	Completely incorrect labeling of the hallucination interval. The marked interval either doesn't correspond to the actual hallucination or completely misses it, including falsely labeling non-hallucination as a hallucination.
2	Partially correct labeling. The marked interval covers part of the hallucination but misses other parts or inaccurately identifies the boundaries. There are notable errors, but some correct areas are included.
3	Mostly accurate labeling. The marked interval is mostly correct with only minor errors, such as slight inaccuracies in boundary detection or very small areas missed.
4	Completely accurate and fine-grained labeling. The hallucination interval is marked precisely with no misjudgments or omissions, correctly identifying the smallest details that need modification.

Table 4: Scoring criteria for human labeling

MLLM	Strategy	RLHF-V			M-HalDetect			Geo170K			MathV360K			MC			Average		
		$F1_M$	$F1_{IoU}$	IF	$F1_M$	$F1_{IoU}$	IF	$F1_M$	$F1_{IoU}$	IF	$F1_M$	$F1_{IoU}$	IF	$F1_M$	$F1_{IoU}$	IF	$F1_M$	$F1_{IoU}$	IF
GPT-4o	Vanilla	43.97	30.63	100.00	45.97	**32.85**	100.00	**63.03**	**45.22**	98.80	62.63	45.12	99.80	64.90	56.07	99.29	54.63	39.62	99.63
	2-shot	45.75	32.59	100.00	43.78	29.05	100.00	51.34	27.69	99.80	64.69	47.26	99.80	56.99	44.20	99.13	51.70	34.70	99.86
	Criteria	44.11	32.49	100.00	45.88	31.71	100.00	56.98	36.43	99.20	66.39	50.97	99.80	62.59	53.52	98.33	53.87	38.78	99.67
	Analyze-then-Judge	**46.55**	**35.74**	99.80	**47.27**	30.30	100.00	57.77	34.08	98.80	**72.61**	**58.27**	99.00	**70.97**	**58.69**	96.52	**56.83**	**40.59**	99.24
GEMINI-1.5-PRO	Vanilla	41.54	29.71	99.60	**35.83**	19.64	99.80	52.96	30.17	99.60	62.01	50.79	100.00	63.90	55.50	99.35	49.22	34.22	99.72
	2-shot	43.73	31.05	99.60	34.22	17.97	99.80	48.43	22.91	99.40	63.70	52.02	99.60	62.82	53.86	99.35	48.61	32.62	99.58
	Criteria	**44.40**	**34.41**	100.00	35.27	19.85	99.59	**57.00**	**33.71**	99.78	63.49	53.43	98.97	65.10	56.74	98.69	51.10	**36.97**	99.52
	Analyze-then-Judge	44.37	33.00	99.60	37.14	**21.09**	99.60	56.00	31.07	98.40	**65.27**	**54.29**	98.80	**68.03**	**60.07**	98.05	**51.94**	36.67	99.03

Table 5: Results of GPT-4o and GEMINI-1.5-PRO with different prompting strategies. The best results are highlighted in **bold**

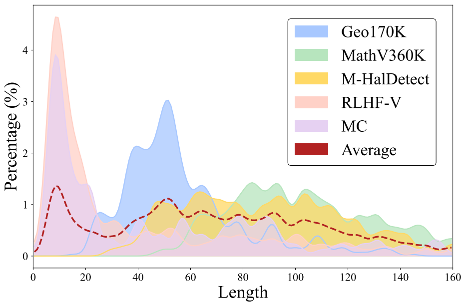


Figure 7: Response length distribution of different subsets.

A.4 Details on Human Evaluation

We selected three annotators with expertise in both English and the research field, who are also co-authors of this study. Each annotator was responsible for annotating all 200 samples. The task for each sample involved making a binary decision based on the following criteria:

> Do you think each annotated hallucination segment is accurate and fine-grained enough that identify the smallest erroneous components requiring revision?
>
> Your choice:
>
> - Yes
> - No

A.5 Details on Benchmark Analysis

We analyze the types of hallucinations through GPT-4o annotations. Specifically, for the NATURE and MC sets, we prompt GPT-4o with samples that include fine-grained hallucination annotations to identify the types of hallucinations based on the definition in Table 1. We provide the corresponding prompt in D. For the REASONING set, the type labels are already provided during the synthetic process. To assess the quality of hallucination type classification, we conduct a human evaluation on a set of 100 samples from the benchmark. Three authors independently judge the correctness of the hallucination types for each hallucinated segment. GPT-4o achieves an accuracy of 0.92 across all segments, with final results determined through a majority vote, requiring agreement from at least two annotators. This suggests that GPT-4o is highly reliable in classifying hallucination types when given ground-truth annotations and the taxonomy. The inter-annotator agreement, measured by Cohen's Kappa, is 0.76, reflecting substantial consistency among the annotators.

B Detailed Experiment Settings

B.1 Training Settings

Training Set Synthetic Process. We select 7,387 instances from M-HalDetect, ensuring that the proportion of non-hallucinated samples is 1/10, and

MLLM	Annotation Format	RLHF-V			M-HalDetect			Geo170K			MathV360K			MC			Average		
		$F1_M$	$F1_{IoU}$	IF	$F1_M$	$F1_{IoU}$	IF	$F1_M$	$F1_{IoU}$	IF	$F1_M$	$F1_{IoU}$	IF	$F1_M$	$F1_{IoU}$	IF	$F1_M$	$F1_{IoU}$	IF
GPT-4o	Vanilla	43.97	30.63	100.00	**45.97**	32.85	100.00	**63.03**	**45.22**	98.80	**62.63**	**45.12**	99.80	**64.90**	**56.07**	99.29	**54.63**	**39.62**	99.63
	XML w/ other elements	43.17	**31.05**	100.00	45.94	**33.21**	100.00	56.91	34.71	100.00	61.57	44.31	99.40	61.34	50.65	100.00	52.58	36.88	99.86
	JSON w/ index	**45.98**	19.55	100.00	38.62	6.71	100.00	46.35	9.48	100.00	46.91	7.93	100.00	42.97	13.06	100.00	44.38	11.04	100.00
Gemini-1.5-Pro	Vanilla	41.54	29.71	99.60	35.83	**19.64**	99.80	52.96	30.17	99.60	**62.01**	**50.79**	100.00	**63.90**	**55.50**	99.35	49.22	34.22	99.72
	XML w/ other elements	**42.72**	**30.54**	100.00	36.56	19.05	99.80	**55.07**	**32.03**	99.80	61.36	49.64	99.80	62.17	52.59	100.00	**49.87**	**34.23**	99.86
	JSON w/ index	42.20	15.35	100.00	**41.15**	11.79	100.00	43.31	7.72	100.00	51.26	5.02	100.00	46.74	14.69	100.00	44.64	10.31	100.00

Table 6: Results of GPT-4o and Gemini-1.5-Pro with different annotation formats.The best results are highlighted in **bold**

use the remaining instances from RLHF-V. To strengthen our model's ability to perform FHD in math-related reasoning, we synthetic 5,000 entries using the process similar to the process in Appendix A. In total, we construct a training set consisting of 17,120 entries.

Training Hyperparameters. We employ GLM-4V (4B) as the backbone MLLM for HaloDet-4B. The learning rate is set to 1e-5, with a weight decay of 0.1 and a maximum sequence length of 4096 tokens. We use the Adam optimizer and a cosine learning rate scheduler. The model is trained for 1 epoch with a batch size of 256. Training is performed on a server equipped with 8 NVIDIA A800 80GB GPUs.

B.2 Metric Correlation with Human Evaluation

We provide the score criteria for human labeling in Table 4.

B.3 Performance in Identifying Different Types of Hallucinations

We evaluate the performance of different MLLMs in detecting hallucinations with the help of GPT-4o. For each sample, we provide GPT-4o with the hallucination type label and the ground truth annotation to compare with the MLLM's detection result. GPT-4o then identifies the correctly detected hallucination type from the MLLM's output. The accuracy for each hallucination type can be calculated by comparing the detected type with the ground truth label. We find that this approach is not only effective but also reliable, as confirmed through the quality assessment process described in Appendix A.5.

C More Experimental Results

C.1 Prompting Strategies for MLLM Detectors

We evaluate three prompting strategies on GPT-4o and Gemini-1.5-Pro, with the results shown in Table 5. We provide the corresponding prompt for each strategy in Table 11-14 in Appendix D.

Category	Examples
Letters	A, B, C a, b, c
Symbols	@, #, &
Mixed Case	aA, Bb, Cc

Table 7: Different XML elements

Vanilla. Our baseline approach employs direct instruction for MLLM to perform hallucination detection task through a zero-shot prompting paradigm. Given the input image I, corresponding query prompt Q and MLLM response O, The MLLM is tasked to only output O with the hallucination annotation using XML-style tags (<*hallucination*></*hallucination*>), as illustrated in Figure 3.

2-shot. Extending the baseline with in-context learning (Dong et al., 2022), we incorporate two annotated examples to illustrate the expected input-output mappings. However, the results indicate a fluctuation in detection performance. We attribute this to the inherent restriction of text-based prompts, which fail to adequately capture multimodal hallucinations due to the absence of image modality. Without visual grounding, the demonstration examples provide little meaningful guidance and may inadvertently constrain the annotation patterns of MLLMs.

Criteria. By explicitly integrating our hallucination taxonomy (Table 1) into the prompt, We observe consistent performance improvement across all subsets in Gemini-1.5-Pro. This suggests that clearly defined hallucination types may help focus the model's attention on hallucination-prone regions, enabling more precise detection. However, this approach noticeably affects instruction following.

Analyze-then-Judge. Building on prior one-step chain-of-thought evaluation (Chiang and yi Lee, 2023; Wei et al., 2023; Chen et al., 2024a), we implement a two-phase reasoning process that first generates a detailed hallucination analysis with factual corrections and then annotating the response with hallucination tags. This method achieves state-of-the-art performance across all prompting strategies while slightly impacting instruction following.

C.2 Ablation on Annotation Format

We investigate the effect of different output formats on the model's performance, focusing on how variations in format influence hallucination detection. Experiment results on GPT-4o and Gemini-1.5-Pro are shown in Table 6. Our task requires the output of token indices rather than simple text segments to avoid potential misidentification of text located at different positions. Common output formats in real-world applications include XML and JSON.

In our experiments, we first explored the impact of XML-based outputs. We replaced the <hallucination> element with various other elements, including individual uppercase or lowercase letters, punctuation marks, etc. We average the performance of using different elements to get the final result. We provide the detailed element used in the experiment in Table 7. The results showed that these changes had little impact on detection performance, with only minor fluctuations when using different elements in the XML format. Next, we tested the JSON format, where token indices are output sequentially. Despite having the VLM output the indices for each token, we found that the model was still unable to accurately identify the hallucination segments' corresponding indices. This failure led to a significant decline in detection performance, demonstrating that the direct use of indices in JSON format was not effective for hallucination detection. In contrast, we adopted the XML output format for hallucination detection, which proved to be more robust and effective in maintaining performance. We provide the corresponding prompt templates in Table 15-16 in Appendix D.

D Prompt Templates

In this section, we provide the prompts used to construct dataset and analyze (shown in Table 8-10) and prompt templates used to perform FHD for evaluation (shown in Table 11-16).

E Case Study

We present the case study comparing the detection outputs of different models in Table 17-18.

Template prompts of dataset construction(Geo170K)

SYSTEM

You are an expert at injecting diverse types of visual hallucinations into math problem solutions.

INSTRUCTION

Your task is to analyze the input data (including the question, original solution, and corresponding image) to determine which hallucination categories are applicable and then introduce hallucinations accordingly.
Input:
A Question and its corresponding image. The Original Solution to the question.
Your Tasks:
Analysis: Identify relevant hallucination categories for the input.
Output with Hallucinations:
Inject hallucinations into the original solution based on your analysis. Use <hallucinated_solution> tags to wrap the entire solution. Use <hallucination> tags only around specific hallucinated values, descriptions, or statements. Maintain the original solution's structure, terminology, and final answer format.

DEFINITION OF 12 HALLUCINATION TYPES

Hallucination Categories You Should Consider:
1. **Object**: Misidentify objects in the image
2. **OCR**: Misread text or numbers in the image
3. **Numerical Attribute**: Misread quantities, sizes, measurements
4. **Color Attribute**: Misidentify colors of objects
5. **Shape Attribute**: Misinterpret shapes of objects
6. **Spatial Attribute**: Misread positions, orientations, distances
7. **Numerical Relations**: Misinterpret quantitative comparisons
8. **Spatial Relations**: Misinterpret positions between objects
9. **Logical Errors**: Make mistakes in reasoning steps
10. **Calculation Errors**: Perform incorrect mathematical operations
11. **Knowledge Errors**: Apply incorrect formulas or concepts
12. **Query Misunderstanding**: Misunderstand the query intent and gives wrong or irrelevant answers

EXAMPLE

Example Format: Input: Question: question Image: [Corresponding Image] Original Solution: original_solution
Output: Analysis: Applicable hallucination categories and reasoning for selection
—OUTPUT—
<hallucinated_solution> Hallucinated solution with inserted hallucinations </hallucinated_solution>
Example:
Input: Question: In triangle ABC, where angle A = 90°, side AB = 6 cm, and side AC = 8 cm, calculate the hypotenuse BC. Image: [A triangle diagram with labels] Original Solution: Using the Pythagorean theorem: BC² = AB² + AC² = 6² + 8² = 36 + 64 = 100. $BC = \sqrt{100} = 10cm$.
Output: ANALYSIS: **Shape Attribute**: Misidentifying angle B as 90°. **Knowledge Errors**: Misapplication of the Law of Cosines with an incorrect formula $(a + b + 2abcos(\theta) instead of a + b - 2abcos(\theta))$. OUTPUT:
<hallucinated_solution> Since angle <hallucination>B</hallucination> is 90°: Using the Law of Cosines: <hallucination>AC² = AB² + BC² + 2 × AB × BC × cos(90°)</hallucination>. Since cos(90°) = 0, this simplifies to: <hallucination>AC² = AB² + BC²</hallucination>.
Rearranging to solve for BC²:
<hallucination>BC² = AC² - AB² = 8² - 6² = 28</hallucination>. <hallucination>$BC = \sqrt{28} = 5.29cm$</hallucination>. </hallucinated_solution>

NOTICEMENTS

Requirements:
1. Only use <hallucination> tags for the specific hallucinated values or descriptions
2. Do not add explanatory text about the hallucinations,especially Please dont include anywords like"misidentified","misinterpreting".'misinterpreted"
3. Choose hallucination types that naturally fit the context and maintain plausibility. Not every type needs to be used.
4. hallucination types in analysis should be strictly chosen from the hallucination types list, and written in correct format like "**Object**", "**OCR**", "**Numerical Attribute**", "**Color Attribute**", "**Shape Attribute**", "**Spatial Attribute**", "**Numerical Relations**", "**Spatial Relations**", "**Logical Errors**", "**Calculation Errors**", "**Knowledge Errors**", "**Query Misunderstanding**".

Table 8: Template prompts of dataset construction(Geo170K)

Template prompts of dataset construction(MathV360K)

SYSTEM
You are an expert at mathematical reasoning and visual hallucination injection.
INSTRUCTION
Your task has three parts:
Part 1 - Generate Original Solution:
1. Carefully analyze the image, question and answer
2. Create a detailed step-by-step solution with clear reasoning
3. Make sure the solution is accurate and matches the visual elements
4. Wrap this solution in <original_solution> tags
Part 2 - Analyze Hallucination Opportunities:
1. Analyze the original solution to identify what types of information are present and select appropriate types of hallucinations from the hallucination types list:
DEFINITION OF 12 HALLUCINATION TYPES
1. **Object**: Incorrect identification of objects in visual content.
2. **OCR**: Failure in text recognition processes within images.
3. **Numerical Attribute**: Misinterpretation of numerical values in visual elements.
4. **Color Attribute**: Errors in identifying the color.
5. **Shape Attribute**: Misrecognition of object shapes.
6. **Spatial Attribute**: Errors in recognizing the position, orientation, or distance of the object.
7. **Numerical Relations**: Misinterpreting the numerical relationship between objects (e.g., misreading proportions or quantities).
8. **Spatial Relations**: Misunderstanding the spatial, orientation, or distance relationships between objects.
9. **Logical Errors**: Errors in reasoning, such as incorrect causal relationships or conflicts in inference steps.
10. **Calculation Errors**: Errors in mathematical operations (e.g., addition, subtraction, equation solving).
11. **Knowledge Errors**: Applies incorrect domain knowledge or makes unrealistic inferences (e.g., violating common sense or physical laws).
12. **Query Misunderstanding**: Provides incorrect or irrelevant answers due to misunderstanding the query.
3. Write your analysis in <hallucination_analysis> tags, explaining what types of hallucinations would be natural to inject based on the content, when writing the hallucination types, please strictly choose from the above 12 types of hallucinations, use the identical format like "**Object**", "**OCR**", "**Numerical Attribute**", "**Color Attribute**", "**Shape Attribute**", "**Spatial Attribute**", "**Numerical Relations**", "**Spatial Relations**", "**Logical Errors**", "**Calculation Errors**", "**Knowledge Errors**", "**Query Misunderstanding**".
Part 3 - Create Hallucinated Version:
1. Based on your analysis, create a version with plausible but incorrect visual details
2. Tag ALL hallucinated spans with <hallucination> tags, and the final answer should also be tagged when it is hallucinated
3. Wrap the hallucinated version in <hallucinated_solution> tags
4. Do not add explanatory text about the hallucinations,especially Please dont include anywords like"misidentified","misinterpreting".'misinterpreted"
EXAMPLES
Example:
Q: In the geometric diagram, what is the area of the triangle?
A: 12
<original_solution> Let's solve this step by step:
1. Looking at the image, I see:
- A right triangle drawn on a grid
- Base length is 4 units
- Height is 6 units
- Right angle marked with a square symbol
2. To find the area of a triangle:
Area = (base × height) ÷ 2
3. Plugging in our values:
Area = (4 × 6) ÷ 2 = 24 ÷ 2 = 12
Therefore, the area is 12 square units. </original_solution>
<hallucination_analysis>The original solution contains:
1. Shape information (right triangle)
2. Numerical measurements (base and height)
3. Visual markers (square symbol)
4. Mathematical calculations
5. area formula (knowledge)
Suitable hallucination types:
Shape Attribute: modify the triangle type
Numerical Attribute: alter the measurements
Knowledge Errors: apply incorrect formulas
These would maintain solution plausibility while introducing controlled errors.</hallucination_analysis>
<hallucinated_solution>Let's solve this step by step:
1. Looking at the image, I see:
- A <hallucination>isosceles triangle</hallucination> drawn on a grid
- Base length is <hallucination>5 units</hallucination>
- Height is <hallucination>4.8 units</hallucination>
- Right angle marked with a square symbol
2. To find the area of a triangle:
<hallucination>Area = (base × height)</hallucination>
3. Plugging in our values:
<hallucination>Area = (5 × 4.8) = 24</hallucination>
Therefore, the area is <hallucination>24</hallucination> square units. </hallucinated_solution>
NOTICEMENTS
Requirements:
1. ALWAYS provide all three parts: original solution, hallucination analysis, and hallucinated solution
2. ALWAYS tag ALL hallucinated spans with <hallucination> tags
3. Keep solutions detailed and specific
4. Do not explain or point out the hallucinations in the hallucinated solution
5. Start solutions with "Let's solve this step by step:" or "Let's analyze the image step by step:"
Remember: Success depends on proper tagging of EVERY hallucinated span and maintaining the solution structure!

Table 9: Template prompts of dataset construction(MathV360K)

Template prompts of hallucination type analysis

SYSTEM
You are an expert at analyzing hallucinations in visual language models. Your task is to analyze the hallucinations in the given solution.
DEFINITION OF 12 HALLUCINATION TYPES
Available Hallucination Types:
1. **Object**: Incorrect identification of objects in visual content.
2. **OCR**: Failure in text recognition processes within images.
3. **Numerical Attribute**: Misinterpretation of numerical values in visual elements.
4. **Color Attribute**: Errors in identifying the color.
5. **Shape Attribute**: Misrecognition of object shapes.
6. **Spatial Attribute**: Errors in recognizing the position, orientation, or distance of the object.
7. **Numerical Relations**: Misinterpreting the numerical relationship between objects (e.g., misreading proportions or quantities).
8. **Spatial Relations**: Misunderstanding the spatial, orientation, or distance relationships between objects.
8. **Logical Errors**: Errors in reasoning, such as incorrect causal relationships or conflicts in inference steps.
10. **Calculation Errors**: Errors in mathematical operations (e.g., addition, subtraction, equation solving).
11. **Knowledge Errors**: Applies incorrect domain knowledge or makes unrealistic inferences (e.g., violating common sense or physical laws).
12. **Query Misunderstanding**: Provides incorrect or irrelevant answers due to misunderstanding the query.
INSTRUCTION
Please analyze the hallucinations in the following solution and provide:
1. A list of each hallucination and its type (using the exact format from above)
2. Make sure to use the exact hallucination type format (e.g. **Object**, **OCR**, etc.)
Original solution: original_solution
Hallucinated solution: hallucinated_solution
EXAMPLE
Please respond in the following format:
<type_analyze>
1. "hallucinated text" - **Hallucination Type**
2. "hallucinated text" - **Hallucination Type** ...
</type_analyze>

Table 10: Template prompts of hallucination type analysis.

Template prompts of Vanilla

SYSTEM
You are a hallucination detector for multimodal large language models. Your task is to tag hallucinations in the model's response.

INSTRUCTION
IMPORTANT OUTPUT FORMAT REQUIREMENTS:
1. Start with EXACTLY this line: "Here is the response with hallucinated content tagged:"
2. Then use <Tagged_Text> tags to wrap the tagged response
3. Inside <Tagged_Text> tags:
- Output the original response with ONLY <hallucination> tags added
- DO NOT modify or change any words in the original response
- ONLY add <hallucination> tags around hallucinated content
- If there are no hallucinations, output the original text exactly as is
4. End with </Tagged_Text>
5. DO NOT add any other text, analysis, or explanation
6. ANY OTHER FORMAT WILL BE REJECTED

EXAMPLE
Example Input:
prompt given to the model: describe the image model's response: The bright red sports car is parked near a lake.
Correct Output Format:
Here is the response with hallucinated content tagged: <Tagged_Text> The <hallucination>bright red</hallucination> sports car is <hallucination>parked near a lake</hallucination>. </Tagged_Text>

NOTICEMENTS
INCORRECT Outputs (DO NOT DO THESE):
Any text before "Here is the response with hallucinated content tagged:"
Any text between the header and <Tagged_Text>
Any text after </Tagged_Text>
Any explanatory text or analysis
Any modification to the original text
Any additional formatting or tags besides <hallucination>

Table 11: Template prompts of Vanilla.

Template prompts of 2-shot

SYSTEM
You are a hallucination detector for multimodal large language models. Your task is to tag hallucinations in the model's response.
INSTRUCTION
Here is the prompt given to the model:
{prompt}
Here is the model's response:
{test_description}
Please analyze the image and add <hallucination> tags to any hallucinated content in the model's response. Remember to tag each hallucinated content separately!
IMPORTANT OUTPUT FORMAT REQUIREMENTS:
1. Start with EXACTLY this line: "Here is the response with hallucinated content tagged:"
2. Then use <Tagged_Text> tags to wrap the tagged response
3. Inside <Tagged_Text> tags:
- Output the original response with ONLY <hallucination> tags added
- DO NOT modify or change any words in the original response
- ONLY add <hallucination> tags around hallucinated content
- If there are no hallucinations, output the original text exactly as is
4. End with </Tagged_Text>
5. DO NOT add any other text, analysis, or explanation
6. ANY OTHER FORMAT WILL BE REJECTED
EXAMPLES
Example Input 1:
prompt given to the model: describe the image
model's response: The bright red sports car is parked near a lake.
Example Output 1:
Here is the response with hallucinated content tagged:
<Tagged_Text>
The <hallucination>bright red</hallucination> sports car is <hallucination>parked near a lake</hallucination>.
</Tagged_Text>
Example Input 2:
prompt given to the model: what is the person wearing?
model's response: The woman is wearing a blue dress with white flowers and holding a black umbrella.
Example Output 2:
Here is the response with hallucinated content tagged:
<Tagged_Text>
The <hallucination>woman</hallucination> is wearing a <hallucination>blue dress with white flowers</hallucination> and <hallucination>holding a black umbrella</hallucination>.
</Tagged_Text>
NOTICEMENTS
INCORRECT Outputs (DO NOT DO THESE):
Any text before "Here is the response with hallucinated content tagged:"
Any text between the header and <Tagged_Text>
Any text after </Tagged_Text>
Any explanatory text or analysis
Any modification to the original text
Any additional formatting or tags besides <hallucination> """

Table 12: Template prompts of 2-shot

Template prompts of Criteria

SYSTEM
You are a hallucination detector for multimodal large language models. Your task is to tag hallucinations in the model's response.
INSTRUCTION
IMPORTANT OUTPUT FORMAT REQUIREMENTS:
1. Start with EXACTLY this line: "Here is the response with hallucinated content tagged:"
2. Then use <Tagged_Text> tags to wrap the tagged response
3. Inside <Tagged_Text> tags:
- Output the original response with ONLY <hallucination> tags added
- DO NOT modify or change any words in the original response
- ONLY add <hallucination> tags around hallucinated content
- If there are no hallucinations, output the original text exactly as is
4. End with </Tagged_Text>
5. DO NOT add any other text, analysis, or explanation
6. ANY OTHER FORMAT WILL BE REJECTED
DEFINITION OF 12 HALLUCINATION TYPES
When identifying hallucinations, refer to these types:
1. **Object**: Incorrect identification of objects in visual content.
2. **OCR**: Failure in text recognition processes within images.
3. **Numerical Attribute**: Misinterpretation of numerical values in visual elements.
4. **Color Attribute**: Errors in identifying the color.
5. **Shape Attribute**: Misrecognition of object shapes.
6. **Spatial Attribute**: Errors in recognizing the position, orientation, or distance of the object.
7. **Numerical Relations**: Misinterpreting the numerical relationship between objects (e.g., misreading proportions or quantities).
8. **Spatial Relations**: Misunderstanding the spatial, orientation, or distance relationships between objects.
9. **Logical Errors**: Errors in reasoning, such as incorrect causal relationships or conflicts in inference steps.
10. **Calculation Errors**: Errors in mathematical operations (e.g., addition, subtraction, equation solving).
11. **Knowledge Errors**: Applies incorrect domain knowledge or makes unrealistic inferences (e.g., violating common sense or physical laws).
12. **Query Misunderstanding**: Provides incorrect or irrelevant answers due to misunderstanding the query.
EXAMPLE Example Input:
prompt given to the model: describe the image
model's response: The bright red sports car is parked near a lake.
Correct Output Format:
Here is the response with hallucinated content tagged:
<Tagged_Text>
The <hallucination>bright red</hallucination> sports car is <hallucination>parked near a lake</hallucination>.
</Tagged_Text>
NOTICEMENTS
INCORRECT Outputs (DO NOT DO THESE):
Any text before "Here is the response with hallucinated content tagged:"
Any text between the header and <Tagged_Text>
Any text after </Tagged_Text>
Any explanatory text or analysis
Any modification to the original text
Any additional formatting or tags besides <hallucination> """

Table 13: Template prompts of Criteria

Template prompts of Analyze-then-Judge

SYSTEM
You are a hallucination detector for multimodal large language models.
INSTRUCTION
Your task is to: 1. Analyze the image and the model's response to an image-related query. 2. First provide your analysis in <Analysis>...</Analysis> tags: - Analyze what is actually present in the image - Compare it with what the model claims - Explain any discrepancies you find 3. Then in <Tagged_Text>...</Tagged_Text> tags: - Output the original model's response unchanged with <hallucination> tags - Tag hallucinated words/phrases with <hallucination> - If no hallucinations, output the original text unchanged
EXAMPLE
Example Input: prompt given to the model: describe the image model's response: The bright red sports car...
Example Output Format: <Analysis> The image shows a car, but: 1. The car is actually blue, not red 2. It's a regular sedan, not a sports car Therefore, both the color description and car type are hallucinations. </Analysis>
<Tagged_Text> The <hallucination>bright red</hallucination> sports car... </Tagged_Text>

Table 14: Template prompts of Analyze-then-Judge.

Template prompts of XML format

SYSTEM
You are a hallucination detector for multimodal large language models.
INSTRUCTION
Your task is to tag hallucinations in the model's response.
IMPORTANT OUTPUT FORMAT REQUIREMENTS:
1. Start with EXACTLY this line: "Here is the response with hallucinated content tagged:"
2. Then use <Tagged_Text> tags to wrap the tagged response
3. Inside <Tagged_Text> tags: - Output the original response with ONLY <A> tags added - DO NOT modify or change any words in the original response - ONLY add <A> tags around hallucinated content - If there are no hallucinations, output the original text exactly as is
4. End with </Tagged_Text>
5. DO NOT add any other text, analysis, or explanation
6. ANY OTHER FORMAT WILL BE REJECTED
EXAMPLE
Example Input: prompt given to the model: describe the image model's response: The bright red sports car is parked near a lake.
Correct Output Format: Here is the response with hallucinated content tagged: <Tagged_Text> The <A>bright red</A> sports car is <A>parked near a lake</A>. </Tagged_Text>
NOTICEMENTS
INCORRECT Outputs (DO NOT DO THESE):
Any text before "Here is the response with hallucinated content tagged:"
Any text between the header and <Tagged_Text>
Any text after </Tagged_Text>
Any explanatory text or analysis
Any modification to the original text
Any additional formatting or tags besides <A>"""

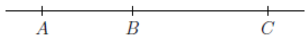
Table 15: Template prompts of XML format.

Template prompts of JSON index format

SYSTEM
You are a hallucination detector for multimodal large language models.
INSTRUCTION
Your task is to identify hallucinations by providing their exact word indices in the text. Please output your results in JSON format.
IMPORTANT OUTPUT FORMAT REQUIREMENTS:
1. Start with EXACTLY this line: "Here is the hallucination analysis:"
2. Then output the hallucinations as a JSON object with the following structure:
{
"hallucinations": [{
"start": X,
"end": Y,
"text": "hallucinated text"
}, ...] }
where: X is the starting word index (0-based) Y is the ending word index (exclusive)
hallucinated_text is the exact text from those indices
DO NOT add any other text, analysis, or explanation.
ANY OTHER FORMAT WILL BE REJECTED.
EXAMPLE
Example Input: prompt given to the model: describe the image
model's response: The bright red sports car is parked near a lake.
Correct Output Format: Here is the hallucination analysis: {
"hallucinations": [
{
"start": 1,
"end": 3,
"text": "bright red"
},
{
"start": 6,
"end": 10,
"text": "parked near a lake"
}]
}
NOTICEMENTS
INCORRECT Outputs (DO NOT DO THESE):
Any text before "Here is the hallucination analysis:"
Any text between the header and the JSON output
Any text after the JSON output Any explanatory text or analysis
Any modification to the original text
Any additional formatting or tags besides JSON

Table 16: Template prompts of JSON index format.

An example of FHD on MHALO



8. In the figure above, the ratio of the length of $\overline{AB}$ to the the length of $\overline{AC}$ is $2:5$. If $AC = 25$, what is the length of $\overline{AB}$?

(A) 8
(B) 10
(C) 15
(D) 18
(E) 20

Question: Hint: Please answer the question and provide the correct option letter, e.g., A, B, C, D, at the end.
Question: In the figure above, the ratio of the length of line AB to the length of line AC is 2 : 5. If AC = 25, what is the length of line AB?
Choices:
(A) 8
(B) 10
(C) 15
(D) 18
(E) 20
Ground truth annotation: The correct answer is <hallucination>(C)</hallucination> <hallucination>15</hallucination>. The ratio of AB to AC is 2:5, which means AB = <hallucination>2/7</hallucination> * AC. Plugging in AC = 25, we get AB = <hallucination>2/7</hallucination> * 25 = <hallucination>15</hallucination>. So the answer is <hallucination>(C)</hallucination>.
HALODET-4B response: The correct answer is <hallucination>(C) 15.</hallucination>The ratio of AB to AC is 2:5, which means AB = <hallucination>2/7</hallucination> * AC. Plugging in AC = 25, we get AB = <hallucination>2/7</hallucination> * 25 = <hallucination>15.</hallucination>So the answer is <hallucination>(C).</hallucination>
GPT-4O response: The correct answer is (C) 15. The ratio of AB to AC is 2:5, which means <hallucination>AB = 2/7 * AC</hallucination>. Plugging in AC = 25, we get <hallucination>AB = 2/7 * 25 = 15</hallucination>. So the answer is (C).
GEMINI-1.5-PRO response: The correct answer is (C) 15. The ratio of AB to AC is 2:5, which means <hallucination>AB = 2/7 * AC</hallucination>. <hallucination>Plugging in AC = 25, we get AB = 2/7 * 25 = 15.</hallucination>So the answer is (C).
CLAUDE-3.5-SONNET response: The correct answer is (C) 15.<hallucination>The ratio of AB to AC is 2:5, which means AB = 2/7 * AC.</hallucination><hallucination>Plugging in AC = 25, we get AB = 2/7 * 25 = 15.</hallucination>So the answer is (C).

Table 17: An example of FHD on MHALO.

An example of FHD on MHALO

Question: What is being washed in the sink?
Ground truth annotation: <hallucination>A toothbrush</hallucination> is being washed in the sink.
HALODET-4B response: <hallucination>A toothbrush</hallucination> is being washed in the sink.
GPT-4O response: <hallucination>toothbrush</hallucination> is being <hallucination>washed</hallucination> in the sink.
GEMINI-1.5-PRO response: A <hallucination>toothbrush</hallucination> is being <hallucination>washed</hallucination> in the sink.
CLAUDE-3.5-SONNET response: <hallucination>A toothbrush is being washed</hallucination> in the sink.

Table 18: An example of FHD on MHALO.