HALLUGUARD: DECOUPLED AUTOMATIC HALLUCINATION DETECTION AND MITIGATION IN MLLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) often exhibit severe hallucinations, where generated text conflicts with visual inputs. Existing evaluation methods heavily rely on costly human annotators or black-box APIs (e.g., GPT-4o). However, these end-to-end evaluators inherently suffer from multimodal entanglement, making them prone to hallucinations themselves and unreliable as judges. To address this, we introduce **HalluGuard**, a decoupled and interpretable framework for hallucination evaluation and mitigation. Instead of relying on black-box MLLMs, our AutoEval module decomposes the complex hallucination detection process into atomic visual perception tasks (e.g., grounding and segmentation). This white-box paradigm provides accurate, fine-grained evaluation for attributes prone to hallucination, such as object quantity, color, and relative position. Based on this deterministic feedback, we further propose Interpretable Hallucination Mitigation (IHM), an inference-time intervention mechanism that effectively corrects hallucinations without requiring any parameter updates or preference data. Furthermore, leveraging our fine-grained metric, we reveal a critical "Goodhart's Law" phenomenon in current MLLM alignment: models often hack hallucination metrics by generating shorter, evasive responses rather than genuinely correcting visual-textual misalignments. Extensive experiments demonstrate that HalluGuard achieves high human agreement in evaluation and effectively mitigates hallucinations, offering a robust alternative to current end-to-end paradigms.

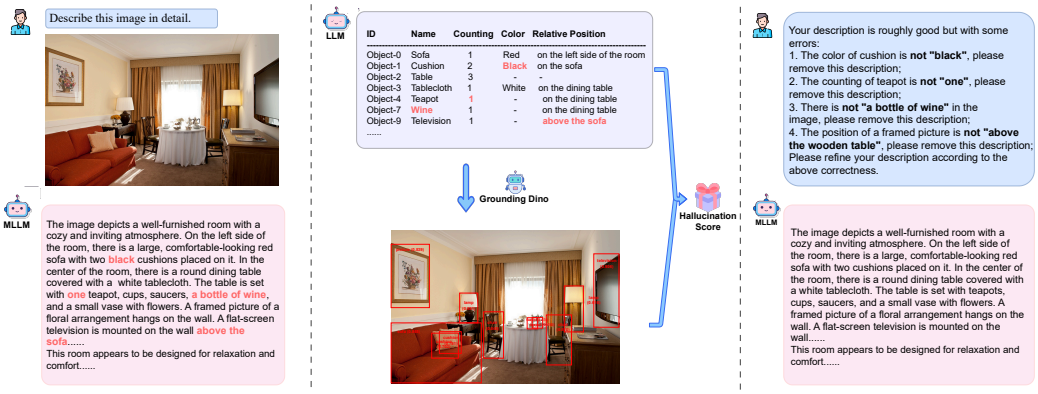
1 INTRODUCTION

The development of Large Language Models (LLMs) (OpenAI, 2022; Google, 2023; Touvron et al., 2023) has offered a glimpse of Artificial General Intelligence (AGI) (Goertzel, 2014; Mitchell, 2024; Xu, 2024). Building upon this foundation, the past few years have witnessed explosive growth in Multimodal Large Language Models (MLLMs) (Jin et al., 2025; Yang et al., 2025b; Li et al., 2023a; Liu et al., 2023b; Li et al., 2024b), which aim to equip LLMs with the ability to "see" the world. However, due to factors such as the imbalance in pretraining across different modalities, the outputs of MLLMs have exhibited varying degrees of hallucinations (Bai et al., 2024). Hallucinations in MLLMs refer to the phenomenon where the textual contents generated by the model conflict with the visual content in the given images (Montagnese et al., 2021), which significantly undermine the accuracy of model responses, hindering their practical applications (Sahoo et al., 2024).

Numerous benchmarks and methods have been developed to detect and reduce hallucinations in MLLMs (Yang et al., 2025a; Sun et al., 2023; Yu et al., 2023). These benchmarks primarily rely on APIs (e.g., GPT-4v, GPT-4o, etc.) (Sun et al., 2023) or expensive human annotators (Yu et al., 2023). However, instead of relying on end-to-end black-box MLLMs which inherently suffer from multimodal entanglement and hallucinations, we argue that a reliable evaluator must be transparent and decoupled. Existing studies Fu et al. (2025) have demonstrated that current multimodal models perform poorly on fundamental visual tasks such as counting and spatial relationships prediction. This inherent vulnerability makes the multimodal models underlying the API prone to generating hallucinations and unsuitable as reliable tools for hallucination detection.

As for hallucination mitigation, a common practice Sun et al. (2023); Yu et al. (2023) is to adapt alignment algorithms (e.g., RLHF, DPO) from those used in Large Language Models (LLMs). The drawbacks of such methods are twofold. First, they necessitate constructing expensive multi-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.



(a) MLLM Output (b) Automatic Evaluation for Hallucination (AutoEval) (c) Interpretable Hallucination Mitigation (IHM)

Figure 1: Illustration of our HalluGuard framework. Best viewed on screen.

modal preference data pairs. Second, they require fine-tuning the model, which brings additional computational cost and potential capability degradation.

To address the issues mentioned above, we propose **HalluGuard**, a decoupled automatic hallucination detection and mitigation framework. HalluGuard consists of two modules: AutoEval (Automatic Evaluation for Hallucination) and IHM (Interpretable Hallucination Mitigation), as illustrated in Figure 1. The core philosophy of AutoEval is *logical decomposition*. Instead of relying on black-box MLLMs, AutoEval decomposes the complex hallucination detection process into atomic visual perception tasks. It provides fine-grained quantitative metrics for specific details including object counting, colors, relative positions, and fabricated objects. Based on this deterministic feedback, IHM reorganizes the extracted hallucination information into a prompt, serving as an inference-time intervention mechanism to instruct the MLLM on how to revise its output without any parameter updates.

Our contributions can be summarized as follows:

- We propose a **decoupled, white-box evaluation paradigm**, AutoEval, which decomposes hallucination detection into atomic visual tasks. It achieves high human correlation and liberates current research from expensive human or inaccurate black-box API evaluation.
- We introduce IHM, an **inference-time intervention mechanism** for hallucination mitigation. It achieves efficient self-correction without the need to construct high-quality preference data or fine-tune the model.
- Leveraging our fine-grained metric, we reveal a critical **"Goodhart's Law" phenomenon** in current MLLM alignment: models often hack the hallucination metrics by generating shorter, evasive responses rather than genuinely correcting visual-textual misalignments. This insight highlights the necessity of our fine-grained metric.

2 HALLUGUARD

By logically decomposing the problem of hallucination detection, we effectively utilize and integrate various existing open-source models to develop an automated hallucination evaluation system that achieves a high level of consistency with human assessments. We first leverage the advanced language capabilities of open-source LLMs to extract a catalog of object names and their corresponding attributes from MLLM responses. Next, we use an existing open-source vision grounding model as a surrogate for human visual inspection, employing the object names in the catalog as queries to retrieve the corresponding bounding boxes and confidence scores. Further, the extracted bounding boxes are fed into an open-source visual color classifier to obtain object colors. If the vision grounding model

does not output any bounding boxes for a given object name, the object is viewed as a hallucination of MLLM.

2.1 HALLUCINATION CATEGORIES

We evaluate four types of fine-grained hallucinations in this work: ***fabrication***, ***counting***, ***color***, and ***relative position*** hallucination, and the ***overall*** hallucination: whether all attributes in textual descriptions are consistent with visual information. According to previous research, fabrication is the most severe and unacceptable type of hallucinations in MLLMs. Due to various causes, MLLMs may describe some objects that do not exist in the given images. Counting hallucination is when MLLMs make mistakes in accurately counting objects. Color hallucination is caused by MLLMs failing to describe the correct colors. Relative position hallucination is that MLLMs wrongly describe the spatial position of one object relative to another.

2.2 AUTOEVAL

As illustrated in Figure 1, AutoEval utilizes open-source LLM and vision grounding models to separately extract information from textual and visual modalities. By determining whether the information corresponding to the same object is consistent across modalities, it detects and measures hallucinations.

Response Parsing Due to the groundbreaking advancements in LLMs, it is easy for various existing LLMs to achieve good performance in text parsing task. To ensure precision, we employ the Qwen-2.5-72B (Team, 2024) to conduct response parsing in our framework. We parse responses of MLLMs into a fine-grained object catalog, including *object names*, *counting*, *colors*, *relative positions*. An example of an object catalog parsed from a description can be found in Figure 1. Prompts for parsing these attributes can be found in Appendix A.

Visual Grounding We use the open-source model, Grounding DINO (Liu et al., 2024), to detect objects from MLLMs' responses. An image I is encoded into visual feature embeddings $F_I = \{f_1, f_2, \ldots, f_N\}$. The text query T is encoded into a textual feature vector $F_T = \phi(T)$. The visual features and the textual feature are aligned using a cross-attention mechanism. For each image feature $f_i \in F_I$, its alignment weight with F_T is computed as:

$$\alpha_i = \text{softmax}\left(\frac{f_i^\top W F_T}{\sqrt{d}}\right), \tag{1}$$

where W is a learnable projection matrix, and d is the feature dimension. Grounding DINO predicts bounding boxes for the relevant image regions. Each bounding box b_k is represented as $b_k = \left(x_k^{\min}, y_k^{\min}, x_k^{\max}, y_k^{\max}\right)$. For each bounding box, the model provides a confidence score $s_k \in [0, 1]$. Failure to ground objects is noted as the first category of hallucinations (fabrication). The number of bounding boxes is used to infer the quantity, and the coordinates are used to infer relative positions.

Visual Color Classification For each detected object, we utilize SAM 2 Ravi et al. (2024) to generate a segmentation mask and replace the colors in the background regions with solid gray. Given the bounding box b_k, SAM 2 outputs a binary segmentation mask $M_k = \text{SAM}_2(I, b_k)$. We then use the visual encoder in CLIP ViT-L/14 Radford et al. (2021b) to extract features from the masked image region, obtaining visual feature v_i. We define a set of colors $C = \{c_1, c_2, \ldots, c_K\}$. For each color c_j, we create a text prompt and encode it to obtain text feature t_j. We compute the cosine similarity between the visual feature vector v_i and the text feature vector t_j:

$$s_{ij} = \frac{v_i \cdot t_j}{\|v_i\|\|t_j\|}. \tag{2}$$

We assign to each object the color that most closely resembles it by finding $\arg\max_j s_{ij}$.

Relative Position Calculation We use the 2D coordinates of the detected bounding box centers to compute relative positions. Let $\mathbf{c_A} = (c_{A_x}, c_{A_y})$ and $\mathbf{c_B} = (c_{B_x}, c_{B_y})$ represent the centers of

objects A and B, where $c_{A_x} = \frac{x_A^{\min} + x_A^{\max}}{2}$, and similarly for others. The offset vector is $\mathbf{o} = \mathbf{c_A} - \mathbf{c_B}$, with magnitude $\|\mathbf{o}\| = \sqrt{\Delta x^2 + \Delta y^2}$. The normalized offsets dx and dy are given by:

$$dx = \frac{\Delta x}{\|\mathbf{o}\|}, \quad dy = \frac{\Delta y}{\|\mathbf{o}\|}. \tag{3}$$

Spatial relationships are determined as follows: If $dx < -0.5$, A is **left of** B; if $dx > 0.5$, A is **right of** B; if $dy < -0.5$, A is **above** B; if $dy > 0.5$, A is **below** B. Relationships can occur in combination (e.g., right and below).

Hallucination Score We design a set of metrics to measure the severity of hallucinations in MLLM responses from five aspects. The first is the overall hallucination metric, which is calculated by determining the percentage of responses containing hallucinations out of the total number of responses generated for a specified image dataset. These responses consist of the detailed descriptions of each image. The remaining four metrics are specifically designed to address four object-related attributes: (1) fabrication of non-existent objects in the image, (2) incorrect object quantity, (3) incorrect object color, and (4) incorrect relative positions between objects. For the first attribute, fabrication, the metric measures the percentage of image descriptions that include objects not present in the actual image, relative to the total number of responses. The calculation of hallucination metrics for the other three attributes follows a similar approach. For the object quantity attribute, since not all image descriptions include specific mentions of object quantities, it is first necessary to identify the subset of responses that do mention object quantities. Within this subset, the metric is calculated as the percentage of responses where the described object quantity does not match the actual object quantity in the image, relative to the total number of responses that include object quantity descriptions. Similarly, the hallucination metrics for the object color and relative positions between objects are calculated in an analogous manner. These metrics assess the percentage of responses containing hallucinations related to object color or relative position inaccuracies out of the total number of relevant descriptions in the dataset.

2.3 IHM

IHM utilizes the hallucination detection results of AutoEval to construct new self-correction instructions, which are fed into the MLLM to guide it in refining its responses. As shown in Figure 1, there are two types of IHM: IHM-Removal and IHM-Correction.

IHM-Removal As shown in the blue background part, IHM-Removal guides the MLLM to directly remove the fine-grained details that are detected as hallucinations. As shown in the example in Figure 1, the MLLM refines its description by removing the hallucinated details after IHM-Removal, such as refining "two black cushions" into "two cushions". IHM-Removal can be applied to the four types of fine-grained hallucinations.

IHM-Correction The IHM-Correction template is shown in the pink background, modifying the hallucinated details into correct information. We only apply IHM-Correction to counting and color hallucinations, and still use IHM-Removal for fabrication and relative position hallucinations. MLLM in the example shown in Figure 1 would correct "two black cushions" into "two orange cushions" if IHM-Correction is applied, using the result of the visual color classifier.

3 BENCHMARKING MLLMS

3.1 SETTINGS

Data Source We randomly select 500 images from COCO 2017 Val set Lin et al. (2014) as our testbed. We prompt various MLLMs to generate detailed descriptions for these images, and quantitatively evaluate hallucinations in the descriptions.

Models We evaluate the latest open-source MLLMs from different model families, including Qwen2.5VL (Team, 2025), Qwen2VL (Wang et al., 2024), InternVL-2.5 (Chen et al., 2024a), InternVL-2 (Chen et al., 2024b), LLaVA-OV (Li et al., 2024a), DeepSeek-VL2 (Wu et al., 2024),

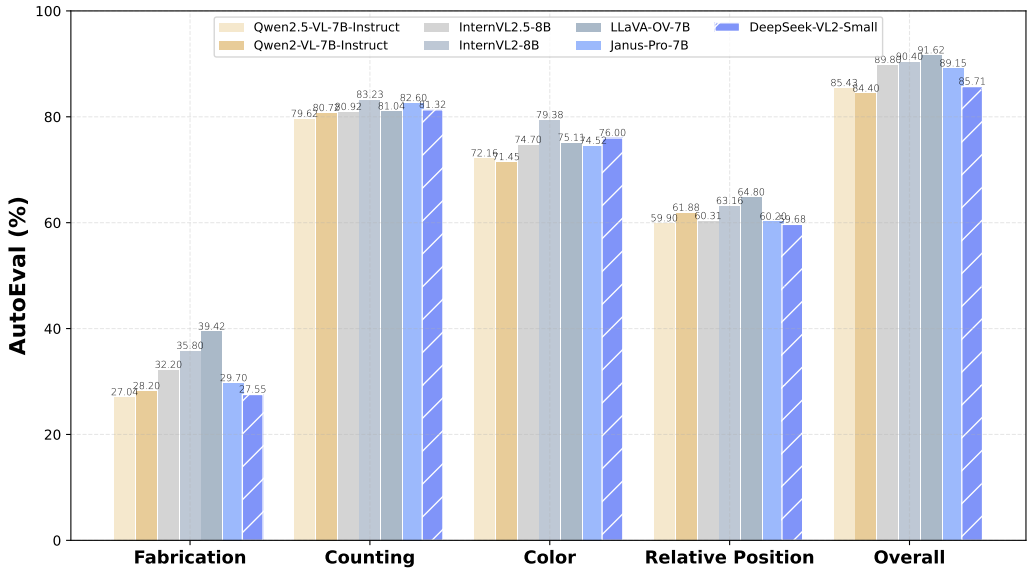


Figure 2: Fine-grained hallucination evaluation results using our AutoEval. A higher score indicates a more severe degree of hallucination.

and Janus-Pro (Chen et al., 2025). We use Qwen-2.5-72B (Team, 2024) and Grounding DINO (Liu et al., 2024) to perform response parsing and visual grounding in our framework respectively. All experiments and model inference are conducted on 8 NVIDIA A100 (80GB) GPU.

Threshold For bounding boxes generated by Grounding DINO, we set the confidence threshold to 0.3 for verification of existence of objects. However, we observe that Grounding DINO tends to pinpoint an excessive number of bounding boxes for object counting, we thus set 0.7 as the confidence threshold for object counting.

3.2 Results

Fine-Grained Evaluation Figure 2 shows the fine-grained hallucination evaluation results. We select models of comparable sizes from different model families for evaluation and comparison. Among the four specified fine-grained hallucination types, all MLLMs exhibit the most severe instances of hallucination in the category of counting. Among all these models, LLaVA-OV-7B has the worst overall performance and the most fabrication hallucinations.

Impact of Model Scale Figure 3 shows the impact of model scale on the overall hallucination for different MLLM families. We observe that the impact of model scale on the AutoEval score is not as significant as that from training schemes between different models. The relationship between model size and the severity of hallucination is not a straightforward linear correlation. For a model family such as DeepSeek-VL2, the overall hallucination score increases with model scale, but for model families such as Qwen2.5-VL-Instruct and Qwen2-VL-Instruct, the overall hallucination score decreases with model scale. Although DeepSeek-VL2-Tiny achieves the lowest hallucination score, the average length of its descriptions is half of that of DeepSeek-VL2-Small, rendering much fewer opportunities for mistakes. The hallucination scores for InternVL2 family are at a relatively high level, with InternVL2-2B having the most severe hallucinations.

Impact of Alignment We separately quantify hallucinations in MLLMs before and after alignment to observe the impact. InternVL2.5-8B uses MPO (Mixed Preference Optimization) for alignment.

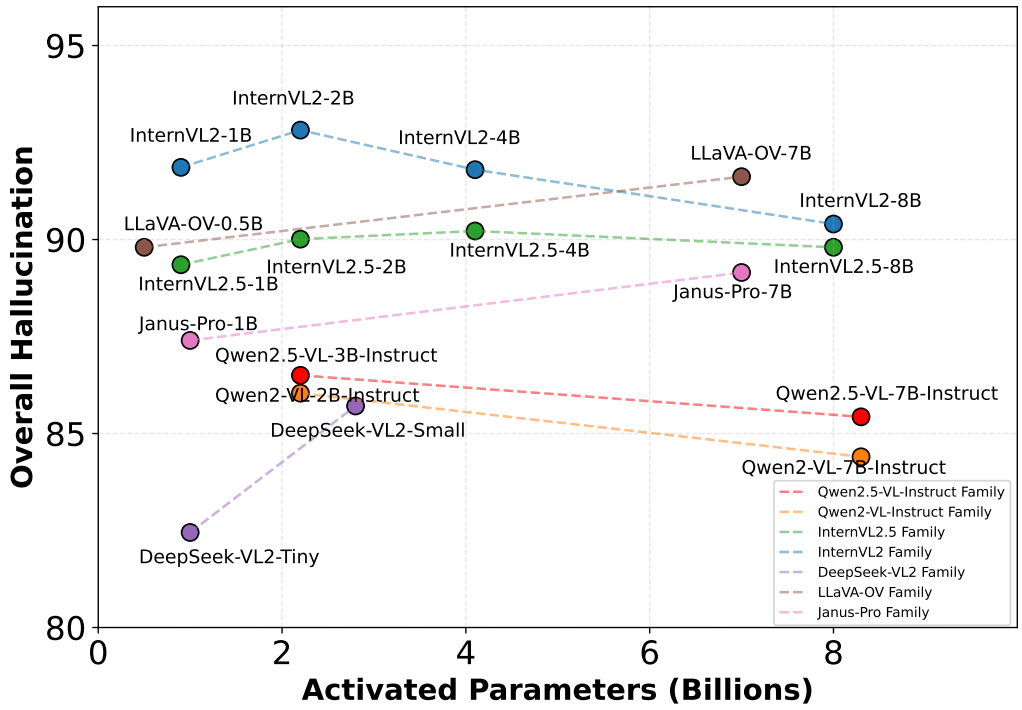


Figure 3: Overall hallucination evaluation across model scale. A higher score indicates a more severe degree of hallucination.

From the results shown in Figure 4, we observe that after alignment training on preference data, hallucination in fabrication, relative position, and color has decreased, but counting hallucination has increased. Although the model's hallucination levels decreases across three dimensions after alignment, we observe that the average length of the descriptions generated by the aligned model shortened from approximately 162 words to about 115 words. This reduction resulted in fewer details being described. We argue that this is hard to say the model has acquired the ability to recognize or correct hallucinations during the alignment process. Instead, it merely indicates that the model has learned to reduce hallucinations by shortening its descriptions.

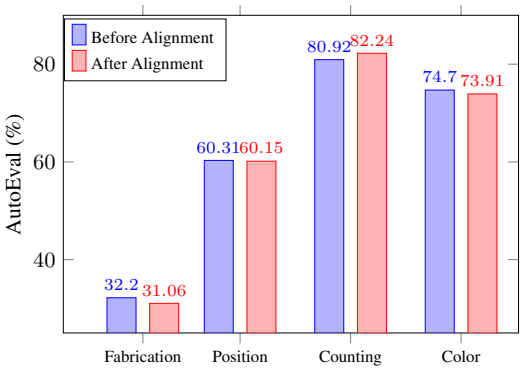


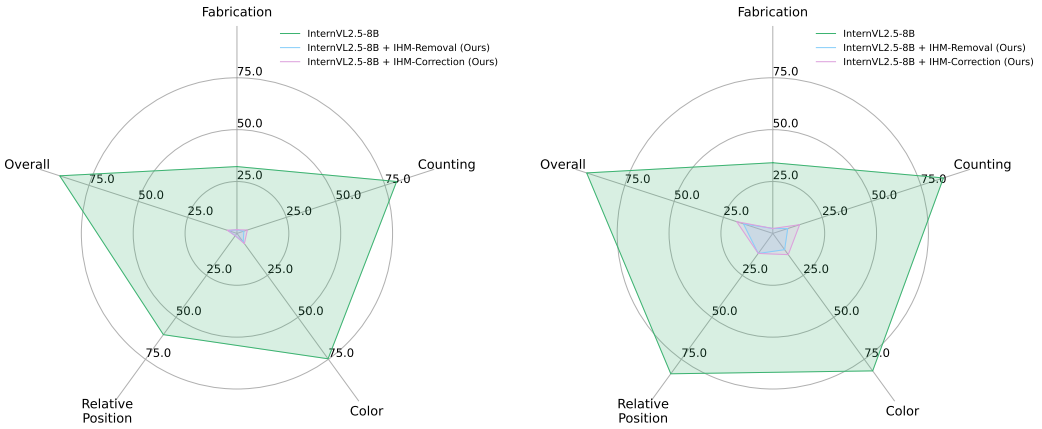
Figure 4: The impact of alignment on fine-grained hallucination. Experiments are conducted on InternVL2.5-8B.

4 Hallucination Mitigation

Based on the interpretable results of AutoEval, we propose a hallucination method IHM, which achieves good performance without additional training.

4.1 Settings

We conduct experiments on InternVL-2.5-8B (Chen et al., 2024a) with two types of IHM, using both AutoEval and human evaluation.



(a) AutoEval evaluation. (b) Human evaluation.

Figure 5: Mitigation results using 2 modes of our IHM on overall and four fine-grained hallucinations.

4.2 Results

Experimental results are shown in Figure 5. IHM-Removal achieves more hallucination mitigation across AutoEval and human evaluation compared with IHM-Correction. Unlike IHM-Removal, which directly removes all erroneous information, IHM-Correction corrects color and counting hallucinations. However, since the visual grounding model and color classifier still exhibit a few errors, there remains a possibility that the corrected counting and color information be identified as hallucinations by human evaluators. In addition, both Removal mode and Correction mode only modify the fine-grained hallucination information detected by our AutoEval, and the modified parts only involve a few words in the description, which has no side effects on the ability and description length of MLLMs.

5 Correlation Study

We conduct a human agreement study of our AutoEval to verify its reliability for hallucination evaluation in MLLMs.

5.1 Settings

We conduct human annotation through Amazon Mechanical Turk. Details of human annotation can be found in Appendix B. We randomly select 100 images from our evaluation set and their corresponding descriptions generated by Qwen2-VL-Instruct, InternVL2, and DeepSeek-VL2-Small, resulting in 300 image-description pairs in total. Three qualified annotators are employed to first list all objects mentioned in the descriptions. Then, for each type of the listed objects, annotators are required to pinpoint their fine-grained attributes (existence, counting, color, and relative position) in the descriptions. These human-annotated object information from descriptions are also used to evaluate the human agreement of the response parsing in our framework. After the annotators extract fine-grained attribute information of objects from descriptions, they are subsequently required to verify whether this information corresponds to the content depicted in the images. Ultimately, based on the preceding annotations of object hallucinations, the annotators provide five binary annotations for the entire description, indicating the presence or absence of fabrication, counting, color, relative position, and overall hallucination.

5.2 Results

Table 1: Human agreement results. "Inter" refers to inter-annotator agreement.

Hallucination	AutoEval	GPT-4o	Inter
Fabrication	0.90	0.74	0.92
Counting	0.81	0.66	0.90
Color	0.80	0.76	0.87
Relative Position	0.81	0.71	0.83
Overall	0.85	0.70	0.88

Table 1 reports the human agreement results for hallucination evaluation. As we can see, our AutoEval achieves 90% agreement with human annotators for overall hallucination evaluation, much higher than that of GPT-4o, which tends to ignore hallucinated details in descriptions and view a hallucinated description as accurate. We find that GPT-4o fails to get a good human agreement, either for overall or fine-grained hallucination evaluation. We speculate that this is primarily because GPT-4o only performs alignment verification on the scenes and main objects across the two modalities, without conducting a detailed, item-by-item examination of the various details that frequently exhibit hallucinations. In fact, with the rapid advancement of MLLM performance, existing MLLMs have almost eliminated errors in describing scenes and main objects, yet they still cannot guarantee the accuracy of all details, which renders challenges in hallucination detection and evaluation. To this end, our method conducts a detailed, item-by-item verification of all objects and their corresponding attribute details across both modalities, thereby enabling the precise detection of even the most subtle hallucinations. The study results also indirectly demonstrate that the open-source modules we employ (response parsing, visual grounding, and visual color classification) exhibit a high degree of human agreement in their respective modality perception tasks. In addition, the high level of agreement among annotators also verifies the validity of our human study.

6 INSIGHTS

Insight 1: MLLMs still suffer from severe hallucinations. As shown in Figure 2, the latest cutting-edge MLLMs still suffer from severe hallucinations and even alignment training has failed to bring about any improvement in this regard. The high level of hallucinations in MLLMs reflects their inaccuracy and unreliability, posing significant obstacles to their practical applications in real-world scenarios. On the eve of Artificial General Intelligence (AGI) (Goertzel, 2014), we hope to see better solutions to avoid unforeseen losses in real-world deployments.

Table 2: Effects of halving description length on hallucination.

MLLM	Fabrication	Count	Color	Position	Overall
Qwen2-VL-2B-Instruct	30.12	77.48	75.20	60.92	82.80
HL Qwen2-VL-2B-Instruct	21.20	73.63	71.32	57.79	74.60
Qwen2-VL-7B-Instruct	28.20	80.72	71.45	61.88	84.40
HL Qwen2-VL-7B-Instruct	22.20	76.72	69.45	56.88	78.05

Insight 2: Hallucinated MLLMs may be more "helpful". We observe that models prone to higher levels of hallucination often provide more detailed descriptions, such as Janus-Pro-7B, compared to models with lower hallucination rates. Based on our preliminary analysis, the number of object categories extracted by these models is twice or even more than that of other models. The increased level of detail in descriptions consequently leads to a higher number of errors. By employing our proposed IHM, it is possible to mitigate hallucinations while preserving as many details as possible.

Insight 3: Beware of Goodhart's Law in hallucination mitigation. *"When a measure becomes a target, it ceases to be a good measure."* (Goodhart, 1984)

If the objective of optimizing MLLMs is restricted to reducing the degree of hallucinations in responses, this may lead to increasingly shorter responses and an avoidance of describing details in the images. After all, if one does not attempt a task, he would never fail. We have conducted experiments on the Qwen2-VL models to study the effects of halving description length on hallucination. Results are presented in Table 2, which shows that shortening descriptions leads to a significant hallucination decrease on all categories. If shortening descriptions provides a straightforward shortcut, why should MLLMs put effort into learning the more challenging task of hallucination correction? In fact, pursuing a low hallucination score is not an optimal goal of MLLM optimization, nor does

it genuinely contribute to MLLMs' advancement. An optimal objective should be to ensure the accuracy of descriptions while providing comprehensive details about the images. We hope to bring attention to Goodhart's law (Manheim & Garrabrant, 2019; Hennessy & Goodhart, 2023) in the hallucination mitigation in MLLMs, instead of blindly pursuing the optimization of hallucination score, since it may cease to be a good measure when MLLMs learn to find a shortcut or hack the reward (Eisenstein et al., 2023).

7 Related Work

Multimodal Large Language Model Multimodal Large Language Models (MLLMs) (Jin et al., 2025; Yang et al., 2025b; Wu et al., 2024; Team, 2025) typically refers to a series of models that enable LLMs to perceive and comprehend data from various modalities. Among them, vision combining LLMs is particularly prominent, owing to the extensive research on vision-language models (VLMs) (Radford et al., 2021a; Li et al., 2022; 2023a). As a result, MLLMs can also be referred to as vision-LLMs (VLLMs) or large vision language models (LVLMs). The goal of MLLMs is to activate the visual capabilities of LLMs, enabling them to "see" the world via images or videos. Combined with strong reasoning and language generation abilities, MLLMs trigger a series of downstream tasks in multimodal domains, such as image or video captioning (Nguyen et al., 2024; Abdar et al., 2024) and visual question answering (Ishmam et al., 2024). Additionally, MLLMs serve as the foundation for applications in other fields, such as AI assistants (Lu et al., 2024), embodied agents (Zhao et al., 2024), and robotics (Sermanet et al., 2024).

Hallucination Evaluation Various benchmarks and metrics have been proposed to assess the extent of hallucinations in the responses of cutting-edge MLLMs (Cai et al., 2025). CHAIR (Rohrbach et al., 2018), POPE (Li et al., 2023c), and MME (Yin et al., 2023) rely on human annotations for objects as ground truth when computing hallucination scores. CIEM (Hu et al., 2023), MMHal-Bench (Sun et al., 2023), GAVIE (Liu et al., 2023a), HaELM (Wang et al., 2023) count on advanced ChatGPT or GPT-4 models to quantify hallucinations in MLLM responses. Unlike these approaches, our AutoEval eliminates the reliance on human annotation and APIs, utilizing existing open-source models to achieve automatic evaluation of hallucinations in MLLMs.

Hallucination Mitigation Alignment methods in LLMs have been introduced to fine-tune MLLMs for hallucination mitigation (Chang et al., 2026). HA-DPO (Zhao et al., 2023) addresses hallucination as a preference selection problem by training models to prioritize accurate responses over hallucinatory ones. Silkie (Li et al., 2023b) first use advanced GPT-4v to label a preference dataset, VLFeedback, and then fine-tune other MLLMs with this dataset through direct preference optimization. POVID Zhou et al. (2024) proposes Preference Optimization in VLLM with AI-Generated Dispreferences (POVID) framework, aiming to exclusively generate dispreferred feedback data using AI models LLaVA-RLHF (Sun et al., 2023) and RLHF-V (Yu et al., 2023) both employ the RLHF paradigm to mitigate hallucinations. Unlike these methods, our IHM achieves good performance in hallucination mitigation without the need for preference data construction and fine-tuning.

8 Conclusion and Limitations

In this paper, we introduced HalluGuard, a decoupled and interpretable framework designed to address the critical challenge of hallucination evaluation and mitigation in MLLMs. By shifting away from entangled, black-box end-to-end evaluators, our AutoEval module decomposes hallucination detection into atomic visual perception tasks, achieving high human agreement across fine-grained attributes. Building on this deterministic feedback, our proposed IHM offers an efficient, training-free inference-time intervention to correct hallucinations. Furthermore, our analysis reveals a concerning "Goodhart's Law" phenomenon in current alignment practices, emphasizing the need for transparent and detailed evaluation metrics. Despite these advantages, HalluGuard currently focuses on object-centric attributes (e.g., fabrication, quantity, color, position), leaving attributes like material, texture, and complex human actions underexplored. Additionally, color recognition is bounded by the predefined categories of the visual classifier. However, owing to the decoupled and modular nature of our framework, future work can seamlessly enhance these capabilities by integrating more advanced atomic perception models. Ultimately, we hope this work inspires the community to rethink MLLM evaluation paradigms and prioritize interpretable, white-box systems.

AI USE STATEMENT

In this work, we used generative AI tools solely to polish the writing, at the level of grammar and phrasing. We did not use generative AI for research ideation, method design, experiments, data collection or analysis, code, or the creation of figures. All AI-polished text was reviewed by the authors, and we take responsibility for the final content of this work, including all text, claims, and artifacts.

ETHICS STATEMENT

This work studies hallucination evaluation and mitigation in publicly available multimodal models on public data (COCO). Human annotation was conducted through Amazon Mechanical Turk with qualified annotators paid an estimated $18 per hour, as detailed in Appendix B. We see no immediate ethical concerns beyond those common to MLLM research.

REPRODUCIBILITY STATEMENT

All components of the framework are built from publicly available models (Qwen-2.5-72B, Grounding DINO, SAM 2, CLIP ViT-L/14), with thresholds specified in the main text and parsing prompts provided in Appendix A. The evaluation set consists of 500 images randomly selected from the COCO 2017 validation set.

REFERENCES

Moloud Abdar, Meenakshi Kollati, Swaraja Kuraparthi, Farhad Pourpanah, Daniel McDuff, Mohammad Ghavamzadeh, Shuicheng Yan, Abduallah Mohamed, Abbas Khosravi, Erik Cambria, et al. A review of deep learning for video captioning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

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

Yishuo Cai, Renjie Gu, Jiaxu Li, Xuancheng Huang, Junzhe Chen, Xiaotao Gu, and Minlie Huang. Mhalo: Evaluating mllms as fine-grained hallucination detectors. In *Findings of the Association for Computational Linguistics: ACL 2025*, pp. 9197–9222, 2025.

Kai-Po Chang, Wei-Yuan Cheng, Chi-Pin Huang, Fu-En Yang, and Yu-Chiang Frank Wang. Mitigating object and action hallucinations in multimodal llms via self-augmented contrastive alignment. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pp. 3172–3181, 2026.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling, 2025.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024a.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024b.

Jacob Eisenstein, Chirag Nagpal, Alekh Agarwal, Ahmad Beirami, Alex D'Amour, DJ Dvijotham, Adam Fisch, Katherine Heller, Stephen Pfohl, Deepak Ramachandran, et al. Helping or herding? reward model ensembles mitigate but do not eliminate reward hacking. *arXiv preprint arXiv:2312.09244*, 2023.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pp. 148–166. Springer, 2025.

Ben Goertzel. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1, 2014.

C. A. E. Goodhart. *Problems of Monetary Management: The UK Experience*, pp. 91–121. Macmillan Education UK, London, 1984. ISBN 978-1-349-17295-5. doi: 10.1007/978-1-349-17295-5_4. URL https://doi.org/10.1007/978-1-349-17295-5_4.

Google. Bard., 2023. URL https://bard.google.com/.

Christopher A Hennessy and Charles AE Goodhart. Goodhart's law and machine learning: a structural perspective. *International Economic Review*, 64(3):1075–1086, 2023.

Hongyu Hu, Jiyuan Zhang, Minyi Zhao, and Zhenbang Sun. Ciem: Contrastive instruction evaluation method for better instruction tuning. *arXiv preprint arXiv:2309.02301*, 2023.

Md Farhan Ishmam, Md Sakib Hossain Shovon, Muhammad Firoz Mridha, and Nilanjan Dey. From image to language: A critical analysis of visual question answering (vqa) approaches, challenges, and opportunities. *Information Fusion*, pp. 102270, 2024.

Yizhang Jin, Jian Li, Tianjun Gu, Yexin Liu, Bo Zhao, Jinxiang Lai, Zhenye Gan, Yabiao Wang, Chengjie Wang, Xin Tan, et al. Efficient multimodal large language models: A survey. *Visual Intelligence*, 3(1):27, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024b. URL https://arxiv.org/abs/2408.03326.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL https://arxiv.org/abs/2201.12086.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023a.

Lei Li, Zhihui Xie, Mukai Li, Shunian Chen, Peiyi Wang, Liang Chen, Yazheng Yang, Benyou Wang, and Lingpeng Kong. Silkie: Preference distillation for large visual language models. *arXiv preprint arXiv:2312.10665*, 2023b.

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European Conference on Computer Vision*. Springer, 2014.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv preprint arXiv:2306.14565*, 1, 2023a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv preprint arXiv:2304.08485*, 2023b.

Shilong Liu, Zhaoyang Zeng, Tianhe Ren, Feng Li, Hao Zhang, Jie Yang, Qing Jiang, Chunyuan Li, Jianwei Yang, Hang Su, Jun Zhu, and Lei Zhang. Grounding dino: Marrying dino with grounded pre-training for open-set object detection, 2024. URL https://arxiv.org/abs/2303.05499.

Ming Y Lu, Bowen Chen, Drew FK Williamson, Richard J Chen, Melissa Zhao, Aaron K Chow, Kenji Ikemura, Ahrong Kim, Dimitra Pouli, Ankush Patel, et al. A multimodal generative ai copilot for human pathology. *Nature*, 634(8033):466–473, 2024.

David Manheim and Scott Garrabrant. Categorizing variants of goodhart's law, 2019.

Melanie Mitchell. Debates on the nature of artificial general intelligence, 2024.

Marcella Montagnese, Pantelis Leptourgos, Charles Fernyhough, Flavie Waters, Frank Larøi, Renaud Jardri, Simon McCarthy-Jones, Neil Thomas, Rob Dudley, John-Paul Taylor, et al. A review of multimodal hallucinations: categorization, assessment, theoretical perspectives, and clinical recommendations. *Schizophrenia Bulletin*, 47(1):237–248, 2021.

Thao Nguyen, Samir Yitzhak Gadre, Gabriel Ilharco, Sewoong Oh, and Ludwig Schmidt. Improving multimodal datasets with image captioning. *Advances in Neural Information Processing Systems*, 36, 2024.

OpenAI. Introducing chatgpt., 2022. URL https://openai.com/blog/chatgpt.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021a. URL https://arxiv.org/abs/2103.00020.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PMLR, 2021b.

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*, 2024.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

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

Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *2024 IEEE International Conference on Robotics and Automation (ICRA)*, pp. 645–652. IEEE, 2024.

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

Qwen Team. Qwen2.5: A party of foundation models, September 2024. URL https://qwenlm.github.io/blog/qwen2.5/.

Qwen Team. Qwen2.5-vl, January 2025. URL https://qwenlm.github.io/blog/qwen2.5-vl/.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton-Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurélien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. Llama 2: Open foundation and fine-tuned chat models. *ArXiv preprint*, abs/2307.09288, 2023. URL https://arxiv.org/abs/2307.09288.

Junyang Wang, Yiyang Zhou, Guohai Xu, Pengcheng Shi, Chenlin Zhao, Haiyang Xu, Qinghao Ye, Ming Yan, Ji Zhang, Jihua Zhu, et al. Evaluation and analysis of hallucination in large vision-language models. *arXiv preprint arXiv:2308.15126*, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, Zhenda Xie, Yu Wu, Kai Hu, Jiawei Wang, Yaofeng Sun, Yukun Li, Yishi Piao, Kang Guan, Aixin Liu, Xin Xie, Yuxiang You, Kai Dong, Xingkai Yu, Haowei Zhang, Liang Zhao, Yisong Wang, and Chong Ruan. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding, 2024. URL `https://arxiv.org/abs/2412.10302`.

Bowen Xu. What is meant by agi? on the definition of artificial general intelligence. *arXiv preprint arXiv:2404.10731*, 2024.

Borui Yang, Md Afif Al Mamun, Jie M Zhang, and Gias Uddin. Hallucination detection in large language models with metamorphic relations. *Proceedings of the ACM on Software Engineering*, 2 (FSE):425–445, 2025a.

Jihan Yang, Shusheng Yang, Anjali W Gupta, Rilyn Han, Li Fei-Fei, and Saining Xie. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 10632–10643, 2025b.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *arXiv preprint arXiv:2306.13549*, 2023.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. *arXiv preprint arXiv:2312.00849*, 2023.

Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023.

Zhonghan Zhao, Wenhao Chai, Xuan Wang, Boyi Li, Shengyu Hao, Shidong Cao, Tian Ye, and Gaoang Wang. See and think: Embodied agent in virtual environment. In *European Conference on Computer Vision*, pp. 187–204. Springer, 2024.

Yiyang Zhou, Chenhang Cui, Rafael Rafailov, Chelsea Finn, and Huaxiu Yao. Aligning modalities in vision large language models via preference fine-tuning. *arXiv preprint arXiv:2402.11411*, 2024.

A Prompt LLM for Response Parsing

Figure 6 shows the prompt for LLM to parse fine-grained object information from descriptions.

B Human Annotation

Our human annotation is conducted through Amazon Mechanical Turk. We employ 3 annotators to annotate 100 samples randomly selected from our evaluation set, and eventually get 3 annotations for each data sample. Every annotator is a native English speaker, who is paid an estimated $18 per hour. Due to the varying length of descriptions and the difference in complexity of images, annotating a single data sample may requires between 1 to 4 minutes. Every annotator goes through a qualification test of 2 hours and is tested to be highly qualified. Specific instructions for annotators are listed below and an example is presented in Figure 7.

Here is a description of an image:

[Description]

Please parse the given description and accurately list all the objects and their corresponding attributes (count, color, relative position) in the following format:

```
Object_ID | Object_Name | Count | Color  | Relative_Position
Object_0  | Apple       | 1     | Red    | on table
Object_1  | Table       | 1     | Yellow | below apple
```

If some attributes are not described, fill in the corresponding slots with '-';
If there are multiple objects of the same type with different attributes (color or relative position), such as a red apple and a green apple mentioned in the description, please place each object on a separate line;
Only when multiple objects of the same type have the same color and relative position, are they merged into the same row.

Figure 6: Prompt for LLM to parse fine-grained object information from descriptions.

Instructions

- Complete question 1 before moving on to other questions. If you are unable to recognize a particular object, guess what it is or use a brief description.
- Additional questions may appear based on your previous answers. Make sure to answer all questions shown without instructions to skip.
- When asked about position, an object can be both above/below and to the left/right of another object. Mark position from your perspective looking at the 2D image, not from the camera perspective.
- Please answer with care: Annotations will be checked by hand according to inter agreement, and work may be rejected if there are too many errors.

Annotation

Description:
The image depicts a group of people sitting outdoors in a relaxed setting, likely a garden or patio area. They are gathered around a small, round glass table, engaged in conversation. Most of the individuals are holding wine glasses, suggesting they are enjoying a social gathering or a wine tasting event. The group consists of both men and women, with some wearing sunglasses and casual summer attire. The setting includes several chairs and tables, some of which are shaded by an umbrella. The background features a well-maintained hedge and some potted plants, indicating a pleasant, outdoor environment. The overall atmosphere appears to be casual and convivial, with people enjoying each other's company and the pleasant weather.



Task 1: Please list all the objects you can find in the given image description. (Separate by commas. If there are no objects in the description, write None.)

Table; chair; ...

Question 2a: Does table exist in the image?

Yes

No

Question 2b-1: What is the quantity of table given in the description? (If this information is not found, please skip to Question c.)

1

Question 2b-2: How many tables are visible in the image? (Either fully or partially.)

2

Question 2b-3: Does the number of tables in the description match that in the image?

Yes

No

Question 2c-1: What is the primary color of table given in the description? (If this information is not found, please skip to Question d.)

Question 2d-1: What is the relative position of table to other objects given in the description? (If this information is not found, please skip to question 3.)

Question 3: Based on the previous annotations, do you think the description contains any hallucinations? (* Hallucination: Anything in the description inconsistent with the image.)

Yes

No

Figure 7: An annotation example in Mechanical Turk.