The Thirty-Ninth AAAI Conference on Artificial Intelligence (AAAI-25)

Combating Multimodal LLM Hallucination via Bottom-Up Holistic Reasoning

Shengqiong Wu[1], Hao Fei[1*], Liangming Pan[2], William Yang Wang[3]
Shuicheng Yan[4,5], Tat-Seng Chua[1]
[1]National University of Singapore, Singapore
[2]University of Arizona, USA
[3]University of California, Santa Barbara, USA
[4]Skywork AI, Singapore
[5]Nanyang Technological University, Singapore
swu@u.nus.edu, {haofei37, dcscts}@nus.edu.sg, liangmingpan@arizona.edu
william@cs.ucsb.edu, shuicheng.yan@gmail.com

Abstract

Recent advancements in multimodal large language models (MLLMs) have shown unprecedented capabilities in advancing various vision-language tasks. However, MLLMs face significant challenges with hallucinations, and misleading outputs that do not align with the input data. While existing efforts are paid to combat MLLM hallucinations, several pivotal challenges are still unsolved. First, while current approaches aggressively focus on addressing errors at the perception level, another important type at the cognition level requiring factual commonsense can be overlooked. In addition, existing methods might fall short in finding a more effective way to represent visual input, which is yet a key bottleneck that triggers visual hallucinations. Moreover, MLLMs can frequently be misled by faulty textual inputs and cause hallucinations, while unfortunately, this type of issue has long been overlooked by existing studies. Inspired by human intuition in handling hallucinations, this paper introduces a novel bottom-up reasoning framework. Our framework systematically addresses potential issues in both visual and textual inputs by verifying and integrating perception-level information with cognition-level commonsense knowledge, ensuring more reliable outputs. Extensive experiments demonstrate significant improvements in multiple hallucination benchmarks after integrating MLLMs with the proposed framework. In-depth analyses reveal the great potential of our methods in addressing perception- and cognition-level hallucinations.

1 Introduction

Recent advancements in MLLMs (Liu et al. 2023a; Wang et al. 2023b; Chen et al. 2023) have demonstrated impressive abilities in understanding the semantics of multimodal data and achieving promising results in a variety of tasks such as visual question answering (VQA; Gurari et al. 2018; Anderson et al. 2018), multimodal dialogue (Wu et al. 2023a; Wang, Zhuang, and Wu 2024; Wu et al. 2024c), image captioning (Milewski, Moens, and Calixto 2020; Liu et al. 2023a; Ji et al. 2021), and retrieval (Fang et al. 2024a; Li et al. 2024; Fang et al. 2024b). However, much like the hallucination issues observed in traditional LLMs due to their generative nature, MLLMs also suffer from hallucinations, leading to factual discrepancies between the model's response and factual reality (Tong et al. 2024; Du et al. 2023). This significantly undermines their practical value. Various types of hallucinations have been identified, prompting many research efforts to address this problem. Existing approaches include fine-tuning MLLMs on more robust instruction datasets (Wang et al. 2024; Liu et al. 2024a), proposing vision-aware inference interventions (Favero et al. 2024; Deng, Chen, and Hooi 2024; Fang et al. 2023), and employing a post-correction method where re-

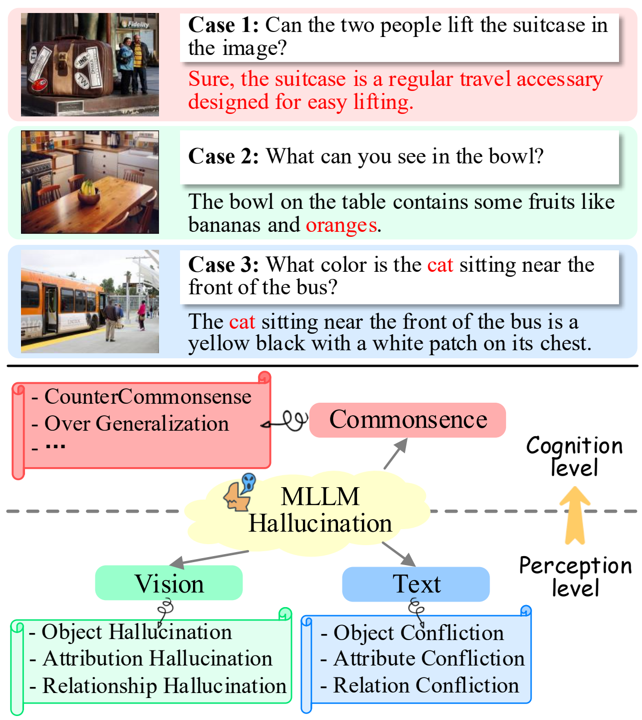

Figure 1: On the top, we illustrate three hallucination cases: overgeneralization (case 1), vision object hallucination (case 2), and text object conflict (case 3), where hallucinations are marked in red. On the bottom, we categorize hallucinations into three types: vision, text, and commonsense, ranging from perception to cognition levels.

*Corresponding author.

sults are generated (Wu et al. 2024a; Yin et al. 2023), then checked and corrected if necessary. Despite these efforts, several critical issues remain unresolved in mitigating hallucinations, especially after deeply revisiting the root-cause of hallucinations from MLLM.

Firstly, a fundamental issue that results in hallucinations in MLLMs is rooted in their insufficient understanding of visual input, which stems from the fact that language-based LLMs do not perceive visual content with the same depth as they understand language. Thus, there is a crucial lack of effective visual representations that assist MLLMs' comprehension of visual data (Fang et al. 2025, 2022; Li et al. 2022a,c). We propose adopting a more structured visual scene representation method, namely Scene Graph (SG), which effectively captures the semantic structure of key objects, attributes, and relationships (Krishna et al. 2017; Wu et al. 2023b; Mitra et al. 2023). This representation naturally aids in addressing the three crucial types of visual hallucinations: object, attribute, and relationship hallucinations (Yang et al. 2024). It is important to note that while current de-hallucination efforts tend to focus on object and attribute hallucinations, typically by employing object detection and phrase grounding tools, often overlooking relationship hallucinations. The introduction of SG representation is expected to resolve these types of issues effectively.

Secondly, existing work often overlooks the fact that input text itself can also be prone to hallucinations. For example, as shown in Figure 1, the input text question, "`What color is the cat sitting near the front of the bus?`" contradicts the image as the input image does not actually present a cat. Such discrepancies may mislead MLLMs, resulting in erroneous outputs that perpetuate hallucinations. Alarmingly, our statistics show that up to 47.8% of user input texts contain such discrepancies, severely impacting performance.

Lastly, upon reevaluating existing methods to address multimodal hallucinations, we observe that most efforts focus on mitigating perception-level hallucinations, such as through object and attribute grounding, but neglect cognition-level checks. Notably, our analysis indicates that issues reliant on cognition-level commonsense knowledge constitute approximately 51% of the problems, particularly in complex input queries. Drawing inspiration from human intuition, we argue that addressing hallucinations should be a holistic reasoning process that incorporates a bottom-up approach from perception to cognition grounding, where perception addresses low-level content awareness and cognition tackles the factuality of commonsense knowledge.

To address the aforementioned challenges, we propose a novel bottom-up reasoning framework for MLLM de-hallucination, from low-level perception grounding to high-level cognition grounding. Inspired by the chain-of-thought (CoT; Mitra et al. 2023; Yao et al. 2023; Fei et al. 2024a, 2023), our framework decomposes the raw visual questions into smaller reasoning subprocesses. These subprocesses progress from high-fidelity perception of target content in both input images and texts to accurately responding with validated cognition-level knowledge. Specifically, our approach can be broken down into six reasoning modules. ①: We guide the MLLM to focus on the visual area most relevant to the user's question, prompting it to generate a partial scene graph. This step ensures the capture of complete visual information, including objects, attributes, and relationships), essential for answering the question. ②: Using external tools, we verify and correct the focused visual content of the partial scene graph representation, ensuring the accuracy of the perceived visual content and preventing hallucinations at the perception level. ③: Based on the high-faithfulness visual perception, we further verify and rectify any discrepancies in the input question that may conflict with the visual content, ensuring accuracy and consistency between the input visuals and text. ④: After confirming the accuracy of both the visual content and the textual question, MLLMs should be able to answer perception-level questions. However, this may not always suffice for cognition-level questions In such cases, the need for additional cognition-level knowledge arises, prompting the MLLM to generate the necessary commonsense claims to answer the question. ⑤: We further verify MLLM-induced commonsense claims against an external knowledge base. ⑥: Integrating all verified perception-level information and cognition-level commonsense, we direct the MLLM to produce the final answer.

We conduct extensive experiments on six benchmarks, demonstrating that the existing MLLMs equipped with our proposed method show significant improvement in mitigating hallucination. In-depth analyses and visualizations show that our method helps decrease conflicts in input questions, thereby reducing erroneous outputs. Overall, our contributions can be summarized into four aspects:

- Drawing inspiration from human reasoning, we propose a novel holistic bottom-up reasoning framework for MLLM de-hallucination, spanning from perception to cognition.
- Our framework innovatively utilizes scene graph representations for visual content during the de-hallucination process of MLLMs.
- We are the first to highlight the often-overlooked issue of hallucinations originating from input text questions, which significantly contribute to hallucinations in user interactions. To address this, we introduce a reconsideration mechanism designed to reduce conflicts between the input queries and visual content, thereby avoiding misleading responses.
- Our proposed framework effectively mitigates various types of hallucination within MLLMs, demonstrating its potential and comprehensiveness.

2 Related Work

Despite the unprecedented capabilities of LLMs, they still exhibit errors on certain NLP tasks, aka., hallucination, due to their generative nature (Liu et al. 2024b). When expanded to MLLMs (Wu et al. 2024b; Fei et al. 2024b; Zhang, Li, and Bing 2023), such an issue persists, characterized by generated text responses that don't align with corresponding visual content (Yin et al. 2023). Research (Yu et al. 2023; Tong et al. 2024; Zhou et al. 2023) categorize hallucinations of MLLM in three types: *object hallucination*, *attribute hallucination*, and *relation hallucination*. To address these, various solutions have been proposed (Zhou et al. 2023; Yin et al. 2023; Liu et al. 2024a). Some suggest refining MLLM with cleaner, more accurate training data (Wang et al. 2024), while others advocate for model intervention during inference (Favero et al. 2024; Deng, Chen, and Hooi 2024), or directly correct model outputs (Wu et al. 2024a; Yin et al. 2023). In this work, we thoroughly rethink the triggers of MLLM hallucination, where certain key aspects that existing works have not fully considered.

On the one hand, MLLMs (Li et al. 2022a) often lack a detailed and accurate understanding of visual images, leading to erroneous outputs. To address this, we propose using scene graph (SG) representations to enhance image comprehension. SGs (Zhang et al. 2024; Cong, Yang, and Rosenhahn 2023; Fei et al. 2024c), highly structured image representations, that precisely capture the semantic meanings of objects, attributes, and their relationships, can intuitively help mitigate all the above three hallucination types by enabling fine-grained and controllable checks of visual faithfulness. On the other hand, aside from the understanding of vision itself, many hallucinations can stem from user textual queries that contain inconsistencies with the visual inputs, misleading MLLM outputs. Unfortunately, this cause has been largely overlooked in prior studies (Wu et al. 2024a; Yin et al. 2023; Favero et al. 2024; Li et al. 2022c).

Last but not least, most research focuses solely on content faithfulness at the recognition level (Wang et al. 2023a; Li et al. 2023; Liu et al. 2024c; Du et al. 2023; Ji et al. 2022), neglecting the factuality checks at the cognitive aspect necessary to prevent hallucinations. Yet this is a critical source for the occurrence of hallucinations, as MLLMs often provide counterfactual responses due to a lack of commonsense (Liu et al. 2024c). Thus, we propose a holistic reasoning framework, inspired by human experts who employ a strict '*from-recognition-to-cognition*' process. We first ensure that all input content (including images and texts) is correctly recognized at the recognition level. Upon building a correct cognition of the inputs, further deep reasoning about the factuality at the cognitive level is conducted.

3 Methodology

Generally, an MLLM, denoted as $f_\theta(\cdot)$ and parameterized by θ, takes the input question Q and visual input I as inputs, conducting reasoning across both modalities: $Y = f_\theta(I, Q)$, where Y is the response. To mitigate the hallucination in MLLMs, we perform a dedicated training-free reasoning framework that decomposes the one-step reasoning process into smaller sub-processes (Xu et al. 2024; Fei et al. 2024a), adhering to the principle of progressing from perception to cognition. The overall structure of the proposed method is illustrated in Figure 2. It consists of six reasoning modules: ① Target Identification and Visual Perception; ② Visual Perception Verification; ③ Question Validation and Adjustment; ④ Commonsense Induction ⑤ Commonsense Verification; ⑥ Question answering. In the following sections, we detail each reasoning module.

Target Identification and Visual Perception. In our first step, we expect to perceive the input text T and vision I by identifying key targets and focusing on visually relevant regions. To accurately represent the perceived visual scene, we employ the SG, denoted $\mathcal{S}_g = \{\mathcal{O}, \mathcal{A}, \mathcal{R}\}$, a structured graph representation of visual scenes, which details not only objects $\mathcal{O}$ within the vision and their corresponding attributes $\mathcal{A}$ but also delineates the relationships $\mathcal{R}$ between objects. Instead of relying on global perception, we prompt the model to infer only from the question what target objects are involved and what partial scene graph can be extracted from the image to answer the question. Therefore, we construct the task prompt $P_{①}$ combined with the input image I to guide the MLLM to generate a partial SG ($\hat{\mathcal{S}}_g$) that is most related to answering the question:

$$\hat{\mathcal{S}}_g = \{\hat{\mathcal{O}}, \hat{\mathcal{A}}, \hat{\mathcal{R}}\} = f_\theta(I, [P_{①}; Q]), \tag{1}$$

where $[;]$ denotes the concatenation operation In practice, we transform the scene graph $\hat{\mathcal{S}}_g$ into JSON format, facilitating an easier interpretation and processing by the MLLMs.

Visual Perception Verification. In the second step, we verify the faithfulness of the partially extracted scene graph from the initial step to ensure accurate subsequent reasoning, concerning that MLLMs are susceptible to hallucinations. We validate the three elements in the scene graph separately. **Firstly**, we use the open-set object detection model, Grounding DINO (Liu et al. 2023b), to verify object fidelity; if the object fails to be detected, we classify the object as low-fidelity and remove it from the scene graph. **Secondly**, for attribute verification, we employ the Grounding DINO model, treating the attributes associated with objects as unified phrases for grounding. **Thirdly**, to ascertain the existence of relationships, we employ the similarity between the image union regions of the subject and object and relationship textual triplets using the BLIP (Li et al. 2022b). By the above three steps, we can ensure the robustness and reliability of the SG for further analysis. Through this module, the verified SG ($\bar{\mathcal{S}}_g = \{\bar{\mathcal{O}}, \bar{\mathcal{A}}, \bar{\mathcal{R}}\}$) will serve as low-level perception evidence (i.e., supporting rationale) for the next process of questioning verification and answering.

Question Validation and Adjustment. In this process, we conduct a further examination of the input question to detect any inconsistencies with the high-fidelity visual perceptions established in the previous process. This verification is crucial; our preliminary experiments reveal that approximately 47.8% of response hallucinations are provoked by pre-existing hallucinations in the input questions, misleading the model into generating content that contradicts the visual facts. For example, Case 3 in Figure 1 demonstrates that the hallucination 'cat' in the question misguided the erroneous generation of responses where the cat is invisible in the image. To prevent such discrepancies from inducing further hallucinations, it is imperative to perform a deconfliction of the question. Specifically, we categorize potential conflicts into three types based on the composition of the visual scene: **1)** `object conflict`, **2)** `attribute conflict`, and **3)** `relationship conflict`. We then employ in-context learning (Wei et al. 2022) to prompt the model to review the input question regarding these three aspects. If conflicts are detected, the model is instructed to

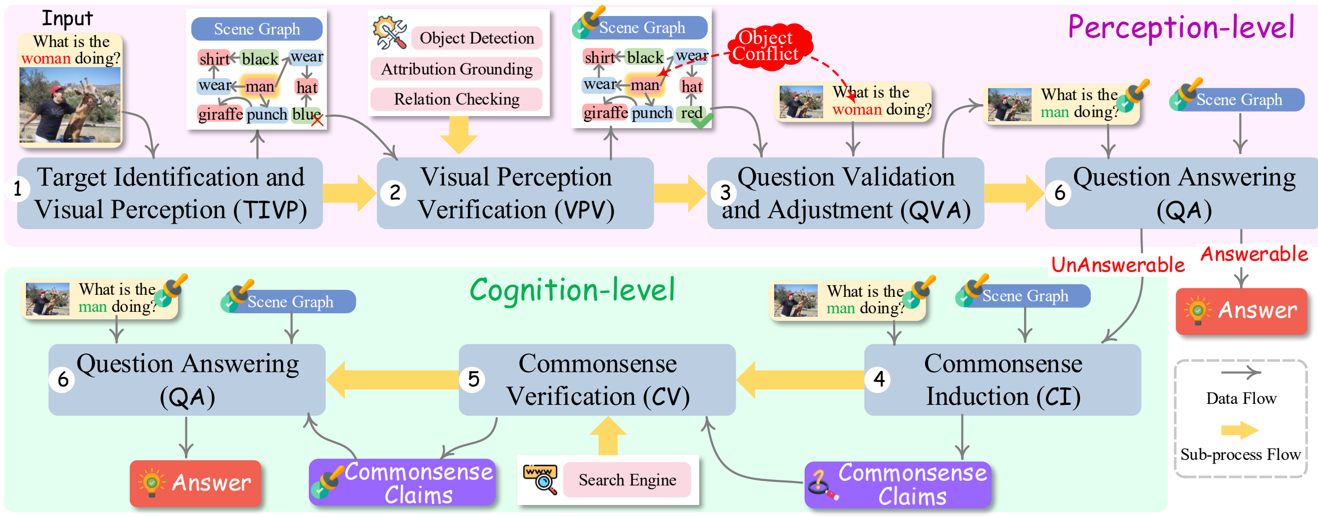

Figure 2: Illustration of the overall framework of DEHALL, consisting of six reasoning modules from perception to cognition.

revise the question to rectify the inconsistency, striving to make minimal semantic alterations to maintain the question's integrity:

$$\bar{Q} = f_\theta(I, [P_{③}; \bar{\mathcal{S}}_g; Q]), \tag{2}$$

where $P_{③}$ is the overall input prompt, and $\bar{Q}$ is the verified input question with a hint of whether the original questions contradict the visual content. In addition, we discover that this module enables the model to engage in active thinking for questions that are unanswerable, ambiguous, or misleading, rather than simply responding with "*I don't know.*"

Commonsense Induction. Having established a comprehensive understanding of the input query and its relevant visual perception, we can now consider answering the question. Although the model shows proficiency in handling perception-level queries based on the verified SG, it still exhibits hallucinations when responding to queries that require cognition-level reasoning, particularly those involving commonsense knowledge. To mitigate this issue, we introduce a commonsense argumentation question-answering mechanism. Specifically, we harness the intrinsic self-analytic capabilities (Wang et al. 2023c) of the LLM to determine whether the question is answerable based on the available evidence. If it is unanswerable, the model outputs the necessary commonsense claims $\hat{\mathcal{C}} = \{\hat{c}_1, \cdot, \hat{c}_n\}$, which forms the fundamental basis required to answer the question:

$$\hat{\mathcal{C}} = f_\theta(I, [P_{④}; \bar{\mathcal{S}}_g; \bar{Q}]), \tag{3}$$

where $P_{④}$ is the task prompt. Through this module, we can obtain the commonsense claims $\hat{\mathcal{C}}$.

Commonsense Verification. This process is designed to validate the faithfulness of the induced `[commonsense claims]`. Technically, we harness the Serper Google Search API to perform web searches using specific fact-based questions. By extracting and scrutinizing the top results, we retrieve a range of fact lists $\mathcal{R}$ from the API's responses for analysis. Then, we leverage the search results to verify the `[commonsense claims]` by prompting the model to categorize each claim $\hat{c}_i$ as either *Hallucination* or *Non-hallucination*:

$$l_i = f_\theta([P_{⑤}; \hat{c}_i; \mathcal{R}]), \tag{4}$$

where $P_{⑤}$ is the overall input prompt for the subprocess-5, and l_i is the label for the claim $\hat{c}_i$. Then, we filter the hallucinated commonsense claims and finally obtain the verified commonsense claims $\bar{\mathcal{C}} = \{\bar{c}_1, \cdot, \bar{c}_m\}$.

Question Answering. Finally, following the aforementioned processes, we have developed a comprehensive bottom-up understanding of the visual elements and questions. Thus, we now prompt the model to provide definitive answers to the question $\bar{Q}$ presented based on the given I and its verified $\bar{S}_g$, and verified commonsense claims $\bar{C}$:

$$Y = f_\theta(I, [P_{⑥}; \bar{S}_g; \bar{C}; \bar{Q}]. \tag{5}$$

Note that the task prompt $P_{⑥}$ slightly varies based on the information available. With only perception-level content, we prompt MLLMs to determine whether it can answer the question. However, after verified commonsense is available, we assume sufficient accurate knowledge is obtained and prompt MLLMs to yield the final answer.

4 Settings

Datasets and Baselines. To rigorously evaluate the performance of the proposed framework, we selected two categories of benchmarks based on the levels at which hallucinations typically occur: `1) Perception-level benchmarks` are used to test the model's ability to dehallucinate visual content concerning objects, attributes, and relationships. This includes benchmarks such as POPE (Li et al. 2023), PHD (Liu et al. 2024c), AMBER (Wang et al. 2023a) and WHOOPS!-VQA (Guetta et al. 2023). `2) Cognition-level benchmarks` are aimed at evaluating the model on more complex issues, such as unanswerable or ambiguous questions, or those requiring common-

Model	OR		AR		SA		PR		C		PhD Avg.		WHOOPS!	
	Neu.	Mis.	Neu.	Mis.	Neu.	Mis.	Neu.	Mis.	Neu.	Mis.	Neu.	Mis.	VQA	Gen.
LLaVA-1.5	65.9	22.5	62.6	11.8	69.0	32.8	47.9	14.5	47.3	11.7	58.5	18.7	47.3	67.9
+ Ours	67.5	35.4	67.0	24.3	76.3	46.5	53.3	29.0	52.1	19.8	63.2 (+4.7)	31.0 (+12.3)	54.5 (+7.2)	72.3 (+4.4)
Qwen-VL-Chat	79.5	46.3	80.9	42.1	73.6	37.9	69.1	43.1	57.6	32.8	72.1	40.4	48.7	67.5
+ Ours	81.8	56.8	86.9	54.0	77.4	47.3	83.4	62.4	64.1	42.3	78.7 (+6.6)	52.5 (+12.1)	54.3 (+5.6)	73.4 (+5.9)
MiniGPT-V2	84.5	43.3	71.5	26.1	78.1	20.5	62.7	35.3	66.1	28.7	72.6	30.8	49.1	71.3
+ Ours	86.0	59.9	76.1	38.0	79.9	42.4	71.0	61.8	68.2	48.8	76.2 (+3.6)	50.2 (+19.4)	51.6 (+2.5)	75.6 (+4.3)
GPT-4V	83.2	76.4	76.2	28.6	76.0	47.2	59.7	42.5	57.6	40.6	70.5	47.1	64.8	81.7
+ Ours	88.0	87.3	85.8	43.7	81.7	65.5	61.8	52.0	85.0	75.4	80.5(+10.0)	64.8 (+17.7)	69.9 (+5.1)	89.8 (+8.1)

Table 1: Evaluation on PhD and WHOOPS! benchmarks. The PhD dataset is split into neural (**Neu.**), and misleading (**Mis.**) questions in Object Recognition (**OR**), Attribute Recognition (**AR**), Sentiment Analysis (**SA**), and Positional Reasoning (**PR**), and Counting (**C**). **PhD Avg.** denotes the average performance on the PhD dataset. For the WHOOPS! benchmark, we evaluate our method on the compositional **VQA** and explanation generation (**Gen.**) tasks.

Model	Acc.	Prec.	Rec.	F1	Yes
LLaVA-1.5*	84.1	90.9	75.7	82.6	41.7
+ Woodpecker$^\flat$	85.7	91.6	76.3	83.2	-
+ LURE†	84.5		-	85.0	-
+ LogicCheckGPT†	90.0	-	-	89.0	-
+ DVP*	85.7	95.5	74.9	84.0	38.4
+ Ours	91.2	96.1	87.4	91.5(+2.5)	43.5
Qwen-VL-Chat*	84.3	94.2	73.0	82.3	38.7
+ Woodpecker$^\flat$	85.7	93.6	76.3	84.0	-
+ LURE$^\flat$	86.7	93.4	79.0	85.5	-
+ LogicCheckGPT$^\flat$	87.6	91.0	83.9	87.3	-
+ DVP*	86.3	99.6	72.8	84.1	36.5
+ Ours	89.7	98.6	86.5	92.2(+4.9)	41.2
GPT-4V*	82.7	85.5	78.8	82.0	46.1
+ Woodpecker$^\flat$	83.1	86.2	79.3	82.6	-
+ LURE$^\flat$	84.4	86.9	81.6	84.2	-
+ LogicCheckGPT$^\flat$	85.9	87.3	83.1	85.1	-
+ DVP*	86.8	88.2	85.0	86.6	48.2
+ Ours	94.2	98.8	89.5	93.9(+7.3)	47.7

Table 2: Performance Evaluation of models on POPE (Adversarial) Setting. the scores with * are derived from (Kim, Kim, and Ro 2024), † are copied from (Wu et al. 2024a), $^\flat$ are re-implemented based on the open-source code.

Model	Uans	FQ	DK	NS	Total
LLaVA-1.5*	4.71	1.62	5.66	0.41	3.73
+ ours	24.67	8.66	10.83	17.69	15.08 (+11.35)
MiniGPT-V2	5.13	1.57	6.70	0.82	3.86
+ ours	25.02	10.62	18.22	19.21	18.97(+15.11)
GPT-4V*	52.02	30.62	49.22	42.21	41.97
+ ours	60.53	33.01	62.85	67.18	55.99(+14.02)

Table 3: Evaluation on VQAv2-IDK dataset. The score with * are copied from (Cha et al. 2024), 'Uans', 'FQ', 'DK', 'NS' denotes 'Unanswerable', 'False Questions', 'Don't Know', 'Not Sure', respectively.

Model	OR	AR	PR	SA	VQA	Gen.
LLaVA-1.5(Ours)	35.4	24.3	46.5	29.0	64.5	74.3
w/o TIVP&VPV	26.6	15.3	33.2	6.8	50.6	68.7
w/o VPV	27.6	17.3	34.3	12.3	58.7	69.1
w/o QAV	24.1	14.1	38.2	15.4	63.0	71.8
w/o CV	31.1	20.4	45.0	25.2	62.8	70.9
w/o CV&CI	30.9	21.8	42.5	18.7	60.3	71.0

Table 4: Ablation study on PhD (misleading questions) and WHOOPS! dataset to validate the efficacy of each module in mitigating hallucination at perception and cognition levels.

sense knowledge, such as sentiment analysis. For this purpose, we selected representative datasets like WHOOPS!-Gen (Guetta et al. 2023) and VQAv2-IDK (Cha et al. 2024). To ensure a comprehensive and impactful assessment, we chose a lineup of representative and widely recognized models, including five Multimodal Large Language Models (MLLMs): LLaVA-1.5 (Liu et al. 2023a), MiniGPT-v2 (Chen et al. 2023), Qwen-VL(Bai et al. 2023), and GPT-4V (OpenAI 2023) (model: gpt-4-vision-preview). We also compare our method with advanced hallucination detection and mitigation methods, including Woodpecker (Yin et al. 2023), LURE (Zhou et al. 2023), LogicCheckGPT (Wu et al. 2024a), and DVP (Kim, Kim, and Ro 2024).

Implementation and Evaluation Metrics. Our framework operates without training, leveraging an open-source pre-trained model to assess performance. We employ Grounding DINO (Liu et al. 2023b) for object and attribute verification and BLIP (Li et al. 2022b) for validating the existence of relationships. To quantify the evaluation, we use accuracy as our metric on the PhD datasets, as followed in (Liu et al. 2024c). For the POPE dataset, we utilize a range of evaluation metrics including Accuracy (Acc.), Precision (Pre.), Recall (Rec.), F1 score (F1), and Yes. Following (Cha et al. 2024), we calculate the IDK Metric to assess performance on the VQAV2-IDK dataset. Additionally, the AMBER (Wang et al. 2023a) score and F1 score are computed to evaluate outcomes on the AMBER and Hal-Eval datasets.

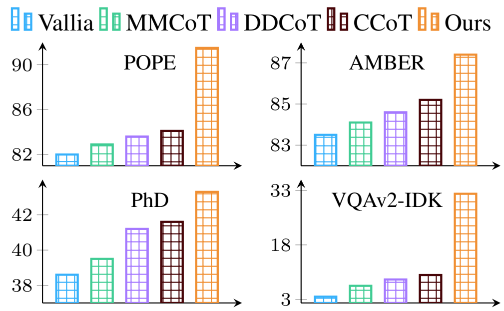

Figure 3: The comparison of different CoT mechanisms.

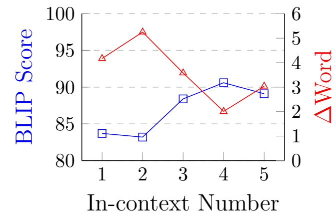

Figure 4: Conflict resolution performance with varying numbers of in-context examples.

5 Experiments and Analysis

Results on Perception Hallucination. We first evaluate our model's performance in mitigating perception-level hallucinations using PhD and POPE datasets. The experimental results, as shown in Tables 1 and 2 reveal a notable performance improvement in our model compared to baseline MLLMs. Specifically, our model achieves an average improvement of 10.8% on the PhD, 4.9% on the POPE, and 5.1% on the WHOOPS!-VQA, with GPT-4V exhibiting a marked enhancement following the integration of our proposed framework. Our model's performance is particularly pronounced when addressing misleading questions, where input questions contain conflicts, compared to neural questions. Moreover, when juxtaposed with baselines designed to mitigate hallucinations, our method consistently displays distinct advantages. Notably, in contrast to the training-dependent LURE model, our approach operates on a training-free basis and still achieves superior performance. Additionally, our method, designed holistically, outperforms traditional post-hoc correction approaches, such as Woodpecker, which typically verify only outputs, leading to notably better performance outcomes.

Results on Cognition Hallucination. Here, we validate the model's ability to mitigate hallucinations at the cognition level. We perform experiments on VQAv2-IDK and WHOOPS!-Gen benchmarks, where the questions require not only perceptual abilities but also cognition-level reasoning skills. As shown in Table 1 and 3, integrating our proposed mechanism leads to a significant performance improvement across various MLLMs, achieving an average of 13.49% improvement on VQAv2-IDK and 6.18% on WHOOPS!-Gen, showcasing the effectiveness of our approach in reducing cognition-level hallucinations.

Ablation Study. To directly assess the contribution of each module in our framework, we conduct an ablation study. The results are detailed in Table 4. Firstly, the removal of any module results in a decline in model performance. Most notably, omitting the visual perception (`TIVP`) and verification (`VPV`) modules lead to the most significant performance deterioration, regardless of whether the questions are at the perception or cognition level. This underscores the importance of accurate perception and verification of image content in vision-language understanding. Furthermore, while eliminating the commonsense induction (`CI`) and verification (`CV`) modules do not significantly impact performance on datasets that primarily focus on perceptual abilities, it significantly affects tasks requiring higher-level reasoning. Specifically, for tasks such as sentiment analysis (SA) or commonsense violation explanation (**Gen.**), which cannot be inferred from visual perception alone, the absence of these modules leads to notable performance degradation.

The Impact of Various CoT. We examine the advantages of our CoT-based framework compared to other CoT frameworks. To this end, we compared our framework against three other CoT-baed methods: MMCoT (Zhang et al. 2023), DDCoT (Zheng et al. 2023), and CCoT (Mitra et al. 2023). As illustrated in Figure 3, it is evident that the design of a CoT framework can significantly reduce model hallucinations. Moreover, while the CCoT framework straightforwardly re-enters an induced scene graph for reasoning, our model incorporates additional subprocesses such as perception verification and input question validation, leading to a significant enhancement in mitigating hallucinations.

Effectiveness of Conflict Resolution in Input question. Next, we examine whether resolving conflicts in input questions enhances reasoning and reduces hallucinations. We conduct experiments on the AMBER dataset, evaluating performance with varying numbers of in-context examples and measuring both the BLIP score (Li et al. 2022b) and the average number of word modifications before and after sentence editing. As shown in Figure 4, our method modifies an average of two words in each input question before and after editing, while maintaining high semantic similarity, indicating that our interventions can de-conflict model inputs without drastically changing the implicit intent comprehension. Furthermore, by visualizing sentences and their corresponding responses both before and after editing in Figure 5, we observe that post-adjusted inputs align more closely with the image, such as minor changes in quantities and objects in images. Most importantly, these modifications ensure that the model's responses are consistent with the visual information provided, effectively mitigating hallucinations.

Case Study. We provide examples to intuitively demonstrate the effectiveness of our proposed framework, as illustrated in Figure 5. Incorporating the proposed framework significantly enhances the proficiency of our model in mitigating hallucinations in input queries and enables it to respond with more detailed rationales. However, as shown in

Case (a)

What color is the **bicycle** in the image?

LLaVA-1.5: The **bicycle** in the image is black.

LLaVA-1.5 + Ours W/O Question Validation: The **bicycle** in the image is black, centrally positioned in the scene and surrounded by other vehicles.

Adjusted Question: What color is the **motorbike** in the image?

LLaVA-1.5 + Ours: There is **no bicycle** in the image you provided; the reference was to a **motorbike**, which is typically larger and mechanically powered compared to a bicycle. If you meant the motorbike, it appears to be black.

Case (b)

What is the three giraffes doing in the image?

LLaVA-1.5: The three giraffes in the image are eating leaves from a tree. They appear to be grazing on the tree, likely enjoying the food it has to offer.

LLaVA-1.5 + Ours W/O Question Validation: In the image, three giraffes are engaging in different activities within a forest. The giraffe in the foreground is standing close to a tree, possibly browsing on the lower branches or the leaves. The other two giraffes appear to be resting.

Adjusted Question: What is the two giraffes doing in the image?

LLaVA-1.5 + Ours: In the image, there are **two giraffes** visible, not three. One of the giraffes is standing and appears to be eating leaves from a tree or possibly browsing on lower branches. The other giraffe is more distant and partially obscured, but it seems to be standing near some trees, possibly also engaging in browsing or simply standing.

Case (c)

Is there a chimpanzee hidden in this image?
A: "No."
B: "Yes, towards the bottom."
C: "Yes, towards the upper middle."
D: "Cannot tell from the image."

Qwen-VL: **D: Cannot tell from the image**. The image shows a dense forest with trees covered in moss and ferns growing on them, but it is not possible to determine if there is a chimpanzee present or not based solely on the image.

LLaVA-1.5: **A: "No."**

Adjusted Question: Is there a chimpanzee hidden in this image?

The scene graph extracted from image:{
"Scene Graphs": {
"objects": [{"id": 1, "type": "Tree", "attributes": {"species": "unknown", "moss": "present", "trunk": "thick", "height": "tall", "location": "forest"}}, {"id": 2, "type": "Tree", "attributes": {"species": "unknown", "moss": "present", "trunk": "thick", "height": "tall", "location": "forest"}}, {"id": 3, "type": "Tree", "attributes": {"species": "unknown", "moss": "present", "trunk": "thick", "height": "tall", "location": "forest"}}, {⋯}]
"relationships": [],
}}

LLaVA-1.5+Ours: **D: "Cannot tell from the image."**

Figure 5: Illustration of example outputs. Case (a) and (b) outputs with and without question validation for input questions containing conflicts. Hallucinations are highlighted in **red** and non-hallucinated in **green**. The input raw questions are marked in green, and the adjusted questions in red. Case (c) shows a failure example.

Figure 5 (c), we observe that some limitations remain, particularly in cases involving ambiguous images where the model's visual information processing capabilities are insufficient, leading to failure in some responses.

6 Conclusion

In this work, we introduce a holistic, training-free reasoning framework to mitigate hallucinations in MLLMs. This framework emulates the human problem-solving process by dividing reasoning into six sub-processes, from perception-level visual understanding to cognition-level commonsense reasoning. Technically, we design a synergistic approach that incorporates perception and cognition-level understanding alongside verification. Additionally, we innovatively propose a question reconsideration and rectification mechanism. Extensive experiments across six benchmarks show that integrating our method into various MLLMs consistently enhances performance on perception and cognition-level questions. Furthermore, in-depth analyses and visualizations reveal that the framework effectively identifies and reduces conflicts between input visual content and questions. The incorporation of verified commonsense further remarkably reduces hallucinations in responses.

References

Anderson, P.; He, X.; Buehler, C.; Teney, D.; Johnson, M.; Gould, S.; and Zhang, L. 2018. Bottom-Up and Top-Down Attention for Image Captioning and Visual Question Answering. In *Proceedings of the CVPR*, 6077–6086.

Bai, J.; Bai, S.; Yang, S.; Wang, S.; Tan, S.; Wang, P.; Lin, J.; Zhou, C.; and Zhou, J. 2023. Qwen-VL: A Frontier Large Vision-Language Model with Versatile Abilities. *CoRR*, abs/2308.12966.

Cha, S.; Lee, J.; Lee, Y.; and Yang, C. 2024. Visually Dehallucinative Instruction Generation: Know What You Don't Know. *CoRR*, abs/2402.09717.

Chen, J.; Zhu, D.; Shen, X.; Li, X.; Liu, Z.; Zhang, P.; Krishnamoorthi, R.; Chandra, V.; Xiong, Y.; and Elhoseiny, M. 2023. MiniGPT-v2: large language model as a unified interface for vision-language multi-task learning. *CoRR*, abs/2310.09478.

Cong, Y.; Yang, M. Y.; and Rosenhahn, B. 2023. RelTR: Relation Transformer for Scene Graph Generation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 45(9): 11169–11183.

Deng, A.; Chen, Z.; and Hooi, B. 2024. Seeing is Believing: Mitigating Hallucination in Large Vision-Language Models via CLIP-Guided Decoding.

Du, L.; Wang, Y.; Xing, X.; Ya, Y.; Li, X.; Jiang, X.; and Fang, X. 2023. Quantifying and Attributing the Hallucination of Large Language Models via Association Analysis. *CoRR*, abs/2309.05217.

Fang, X.; Fang, W.; Liu, D.; Qu, X.; Dong, J.; Zhou, P.; Li, R.; Xu, Z.; Chen, L.; Zheng, P.; et al. 2024a. Not all inputs are valid: Towards open-set video moment retrieval using language. In *Proceedings of the ACM MM*.

Fang, X.; Liu, D.; Fang, W.; Zhou, P.; Xu, Z.; Xu, W.; Chen, J.; and Li, R. 2024b. Fewer Steps, Better Performance: Efficient Cross-Modal Clip Trimming for Video Moment Retrieval Using Language. In *Proceedings of the AAAI*.

Fang, X.; Liu, D.; Zhou, P.; and Hu, Y. 2022. Multi-modal cross-domain alignment network for video moment retrieval. *IEEE TMM*.

Fang, X.; Liu, D.; Zhou, P.; and Nan, G. 2023. You can ground earlier than see: An effective and efficient pipeline for temporal sentence grounding in compressed videos. In *Proceedings of the CVPR*.

Fang, X.; Xiong, Z.; Fang, W.; Qu, X.; Chen, C.; Dong, J.; Tang, K.; Zhou, P.; Cheng, Y.; and Liu, D. 2025. Rethinking weakly-supervised video temporal grounding from a game perspective. In *Proceedings of the ECCV*.

Favero, A.; Zancato, L.; Trager, M.; Choudhary, S.; Perera, P.; Achille, A.; Swaminathan, A.; and Soatto, S. 2024. Multi-Modal Hallucination Control by Visual Information Grounding. *CoRR*, abs/2403.14003.

Fei, H.; Li, B.; Liu, Q.; Bing, L.; Li, F.; and Chua, T.-S. 2023. Reasoning Implicit Sentiment with Chain-of-Thought Prompting. In *Proceedings of ACL*, 1171–1182.

Fei, H.; Wu, S.; Ji, W.; Zhang, H.; Zhang, M.; Lee, M.-L.; and Hsu, W. 2024a. Video-of-thought: Step-by-step video reasoning from perception to cognition. In *Proceedings of the ICML*.

Fei, H.; Wu, S.; Zhang, H.; Chua, T.-S.; and Yan, S. 2024b. VITRON: A Unified Pixel-level Vision LLM for Understanding, Generating, Segmenting, Editing.

Fei, H.; Wu, S.; Zhang, M.; Zhang, M.; Chua, T.-S.; and Yan, S. 2024c. Enhancing video-language representations with structural spatio-temporal alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Guetta, N. B.; Bitton, Y.; Hessel, J.; Schmidt, L.; Elovici, Y.; Stanovsky, G.; and Schwartz, R. 2023. Breaking Common Sense: WHOOPS! A Vision-and-Language Benchmark of Synthetic and Compositional Images. In *Proceedings of the ICCV*, 2616–2627.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the CVPR*, 3608–3617.

Ji, J.; Luo, Y.; Sun, X.; Chen, F.; Luo, G.; Wu, Y.; Gao, Y.; and Ji, R. 2021. Improving image captioning by leveraging intra-and inter-layer global representation in transformer network. In *Proceedings of the AAAI*, 1655–1663.

Ji, J.; Ma, Y.; Sun, X.; Zhou, Y.; Wu, Y.; and Ji, R. 2022. Knowing what to learn: a metric-oriented focal mechanism for image captioning. *IEEE Transactions on Image Processing*, 31: 4321–4335.

Kim, J.; Kim, Y.; and Ro, Y. M. 2024. What if...?: Counterfactual Inception to Mitigate Hallucination Effects in Large Multimodal Models. *CoRR*, abs/2403.13513.

Krishna, R.; Zhu, Y.; Groth, O.; Johnson, J.; Hata, K.; Kravitz, J.; Chen, S.; Kalantidis, Y.; Li, L.; Shamma, D. A.; Bernstein, M. S.; and Fei-Fei, L. 2017. Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations. *Int. J. Comput. Vis.*, 123(1): 32–73.

Li, J.; He, X.; Wei, L.; Qian, L.; Zhu, L.; Xie, L.; Zhuang, Y.; Tian, Q.; and Tang, S. 2022a. Fine-grained semantically aligned vision-language pre-training. *Proceedings of the NeurIPS*, 35: 7290–7303.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. C. H. 2022b. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the ICML*, 12888–12900.

Li, J.; Xie, J.; Qian, L.; Zhu, L.; Tang, S.; Wu, F.; Yang, Y.; Zhuang, Y.; and Wang, X. E. 2022c. Compositional temporal grounding with structured variational cross-graph correspondence learning. In *Proceedings of the CVPR*, 3032–3041.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; and Wen, J. 2023. Evaluating Object Hallucination in Large Vision-Language Models. In *Proceedings of the EMNLP*, 292–305.

Li, Y.; Wang, W.; Qu, L.; Nie, L.; Li, W.; and Chua, T. 2024. Generative Cross-Modal Retrieval: Memorizing Images in Multimodal Language Models for Retrieval and Beyond. In *Proceedings of the ACL*, 11851–11861.

Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; and Wang, L. 2024a. Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2023a. Improved Baselines with Visual Instruction Tuning. *CoRR*, abs/2310.03744.

Liu, H.; Xue, W.; Chen, Y.; Chen, D.; Zhao, X.; Wang, K.; Hou, L.; Li, R.; and Peng, W. 2024b. A Survey on Hallucination in Large Vision-Language Models. *CoRR*, abs/2402.00253.

Liu, J.; Fu, Y.; Xie, R.; Xie, R.; Sun, X.; Lian, F.; Kang, Z.; and Li, X. 2024c. PhD: A Prompted Visual Hallucination Evaluation Dataset. *CoRR*, abs/2403.11116.

Liu, S.; Zeng, Z.; Ren, T.; Li, F.; Zhang, H.; Yang, J.; Li, C.; Yang, J.; Su, H.; Zhu, J.; and Zhang, L. 2023b. Grounding DINO: Marrying DINO with Grounded Pre-Training for Open-Set Object Detection. *CoRR*, abs/2303.05499.

Milewski, V. S. J.; Moens, M.; and Calixto, I. 2020. Are Scene Graphs Good Enough to Improve Image Captioning? In *Proceedings of the AACL*, 504–515.

Mitra, C.; Huang, B.; Darrell, T.; and Herzig, R. 2023. Compositional Chain-of-Thought Prompting for Large Multimodal Models. *CoRR*, abs/2311.17076.

OpenAI. 2023. GPT-4 Technical Report. *CoRR*, abs/2303.08774.

Tong, S.; Liu, Z.; Zhai, Y.; Ma, Y.; LeCun, Y.; and Xie, S. 2024. Eyes Wide Shut? Exploring the Visual Shortcomings of Multimodal LLMs. *CoRR*, abs/2401.06209.

Wang, J.; Wang, Y.; Xu, G.; Zhang, J.; Gu, Y.; Jia, H.; Yan, M.; Zhang, J.; and Sang, J. 2023a. An LLM-free Multi-dimensional Benchmark for MLLMs Hallucination Evaluation. *CoRR*, abs/2311.07397.

Wang, L.; He, J.; Li, S.; Liu, N.; and Lim, E. 2024. Mitigating Fine-Grained Hallucination by Fine-Tuning Large Vision-Language Models with Caption Rewrites. In *Proceedings of the MMM*, 32–45.

Wang, W.; Lv, Q.; Yu, W.; Hong, W.; Qi, J.; Wang, Y.; Ji, J.; Yang, Z.; Zhao, L.; Song, X.; Xu, J.; Xu, B.; Li, J.; Dong, Y.; Ding, M.; and Tang, J. 2023b. CogVLM: Visual Expert for Pretrained Language Models. *CoRR*, abs/2311.03079.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q. V.; Chi, E. H.; Narang, S.; Chowdhery, A.; and Zhou, D. 2023c. Self-Consistency Improves Chain of Thought Reasoning in Language Models. In *Proceedings of the ICLR*.

Wang, X.; Zhuang, B.; and Wu, Q. 2024. ModaVerse: Efficiently Transforming Modalities with LLMs. *arXiv preprint arXiv:2401.06395*.

Wei, J.; Tay, Y.; Bommasani, R.; Raffel, C.; Zoph, B.; Borgeaud, S.; Yogatama, D.; Bosma, M.; Zhou, D.; Metzler, D.; Chi, E. H.; Hashimoto, T.; Vinyals, O.; Liang, P.; Dean, J.; and Fedus, W. 2022. Emergent Abilities of Large Language Models. *Trans. Mach. Learn. Res.*, 2022.

Wu, C.; Yin, S.; Qi, W.; Wang, X.; Tang, Z.; and Duan, N. 2023a. Visual ChatGPT: Talking, Drawing and Editing with Visual Foundation Models. *CoRR*, abs/2303.04671.

Wu, J.; Liu, Q.; Wang, D.; Zhang, J.; Wu, S.; Wang, L.; and Tan, T. 2024a. Logical Closed Loop: Uncovering Object Hallucinations in Large Vision-Language Models. *CoRR*, abs/2402.11622.

Wu, S.; Fei, H.; Cao, Y.; Bing, L.; and Chua, T. 2023b. Information Screening whilst Exploiting! Multimodal Relation Extraction with Feature Denoising and Multimodal Topic Modeling. In *Proceedings of the ACL*, 14734–14751.

Wu, S.; Fei, H.; Li, X.; Ji, J.; Zhang, H.; Chua, T.-S.; and Yan, S. 2024b. Towards Semantic Equivalence of Tokenization in Multimodal LLM. *arXiv preprint arXiv:2406.05127*.

Wu, S.; Fei, H.; Qu, L.; Ji, W.; and Chua, T. 2024c. NExT-GPT: Any-to-Any Multimodal LLM. In *Proceedings of the ICML*.

Xu, J.; Fei, H.; Pan, L.; Liu, Q.; Lee, M.-L.; and Hsu, W. 2024. Faithful Logical Reasoning via Symbolic Chain-of-Thought. *arXiv preprint arXiv:2405.18357*.

Yang, Z.; Feng, R.; Yan, K.; Wang, H.; Wang, Z.; Zhu, S.; Zhang, H.; Xiao, J.; Wu, P.; Zhu, K.; Chen, J.; Xie, C.; Mao, C.; Yang, Y.; Zhang, H.; Liu, Y.; and Cheng, F. 2024. BACON: Supercharge Your VLM with Bag-of-Concept Graph to Mitigate Hallucinations. *CoRR*, abs/2407.03314.

Yao, S.; Zhao, J.; Yu, D.; Du, N.; Shafran, I.; Narasimhan, K. R.; and Cao, Y. 2023. ReAct: Synergizing Reasoning and Acting in Language Models. In *Proceedings of the ICLR*.

Yin, S.; Fu, C.; Zhao, S.; Xu, T.; Wang, H.; Sui, D.; Shen, Y.; Li, K.; Sun, X.; and Chen, E. 2023. Woodpecker: Hallucination Correction for Multimodal Large Language Models. *CoRR*, abs/2310.16045.

Yu, Q.; Li, J.; Wei, L.; Pang, L.; Ye, W.; Qin, B.; Tang, S.; Tian, Q.; and Zhuang, Y. 2023. HalluciDoctor: Mitigating Hallucinatory Toxicity in Visual Instruction Data. *CoRR*, abs/2311.13614.

Zhang, C.; Stepputtis, S.; Campbell, J.; Sycara, K. P.; and Xie, Y. 2024. HiKER-SGG: Hierarchical Knowledge Enhanced Robust Scene Graph Generation. *CoRR*, abs/2403.12033.

Zhang, H.; Li, X.; and Bing, L. 2023. Video-LLaMA: An Instruction-tuned Audio-Visual Language Model for Video Understanding. *CoRR*, abs/2306.02858.

Zhang, Z.; Zhang, A.; Li, M.; Zhao, H.; Karypis, G.; and Smola, A. 2023. Multimodal Chain-of-Thought Reasoning in Language Models. *CoRR*, abs/2302.00923.

Zheng, G.; Yang, B.; Tang, J.; Zhou, H.; and Yang, S. 2023. DDCoT: Duty-Distinct Chain-of-Thought Prompting for Multimodal Reasoning in Language Models. In *Proceedings of the NeurIPS*.

Zhou, Y.; Cui, C.; Yoon, J.; Zhang, L.; Deng, Z.; Finn, C.; Bansal, M.; and Yao, H. 2023. Analyzing and Mitigating Object Hallucination in Large Vision-Language Models. *CoRR*, abs/2310.00754.