

Grounded Chain-of-Thought for Multimodal Large Language Models

Qiong Wu¹, Xiangcong Yang¹, Yiyi Zhou¹, Chenxin Fang¹, Baiyang Song¹, Xiaoshuai Sun¹, Rongrong Ji¹

¹ Key Laboratory of Multimedia Trusted Perception and Efficient Computing,
Ministry of Education of China, Xiamen University, 361005, P.R. China.

{qiong, yangxiangcong}@stu.xmu.edu.cn, zhouyiyi@xmu.edu.cn,

fangchenxin@stu.xmu.edu.cn, songbaiyang07@gmail.com, {xssun, rrji}@xmu.edu.cn

Abstract

Despite great progress, existing multimodal large language models (MLLMs) are prone to visual hallucination, greatly impeding their trustworthy applications. In this paper, we study this problem from the perspective of visual-spatial reasoning, and propose a new learning task for MLLMs, termed Grounded Chain-of-Thought (GCoT). Different from recent visual CoT studies, which focus more on visual knowledge reasoning, GCoT is keen to helping MLLMs to recognize and ground the relevant visual cues step by step, thereby predicting the correct answer with grounding coordinates as the intuitive basis. To facilitate this task, we also carefully design and construct a dataset called multimodal grounded chain-of-thought (MM-GCoT) consisting of 24,022 GCoT examples for 5,033 images. Besides, a comprehensive consistency evaluation system is also introduced, including the metrics of answer accuracy, grounding accuracy and answer-grounding consistency. We further design and conduct a bunch of experiments on 12 advanced MLLMs, and reveal some notable findings: i. most MLLMs performs poorly on the consistency evaluation, indicating obvious visual hallucination; ii., visual hallucination is not directly related to the parameter size and general multimodal performance, i.e., a larger and stronger MLLM is not less affected by this issue. Lastly, we also demonstrate that the proposed dataset can help existing MLLMs to well cultivate their GCoT capability and reduce the inconsistent answering significantly. Moreover, their GCoT can be also generalized to exiting multimodal tasks, such as open-world QA and REC. Our dataset and evaluation scripts are anonymously released at: <https://github.com/DoubtedSteam/MM-GCoT>

1. Introduction

As a research hot-spot, *Multimodal Large Language Model* (MLLM) [1–3, 22, 28, 41] has made remarkable progresses in approaching near-human vision-language (VL) capabil-

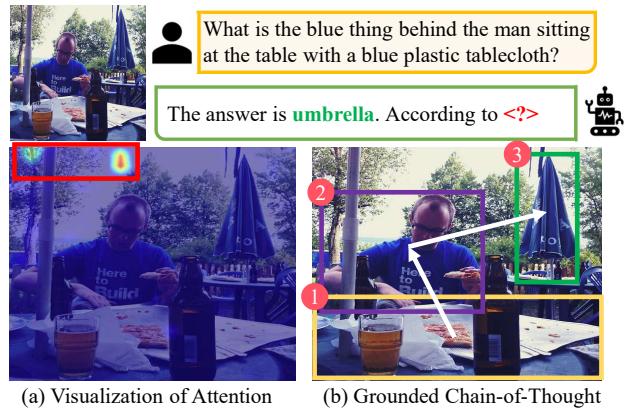


Figure 1. The comparison between common question answering (a) and the proposed *grounded chain-of-thought* (GCoT) (b). (a) An MLLM can directly output the correct answer, which however is often not based on what it, as shown in its attention heat map. This issue undermines its trustworthy application. (b) GCoT aims to help the MLLM make grounded reasoning step-by-step, and outputs the answer with coordinates as the intuitive basis.

ties on various benchmarks [6, 8, 13, 31]. However, visual hallucination [40, 52] still remains a critical problem that long plagues the application of MLLMs. In practice, MLLMs may fail to answer the question due to insufficient understanding of the given image. This visual shortcoming is easy to be measured via existing evaluation systems [9, 15, 24], and it also arouses great attention and receive numerous recent efforts [14, 27, 42, 51, 52].

A more subtle manifestation is when an MLLM can answer the question accurately, but relies more on data distribution bias than key visual information, also known as *language bias* in previous VL study [11, 46, 49]. As shown in Fig.1, although the answer is correct, the MLLM attend to irrelevant regions for answering, so it also naturally ground the wrong answer cue. Although this issue does not affect the performance evaluation of MLLMs on most VL benchmarks, it is easy to misvalue their reliability and bring potential application risks. In this case, some efforts [4, 35, 50]

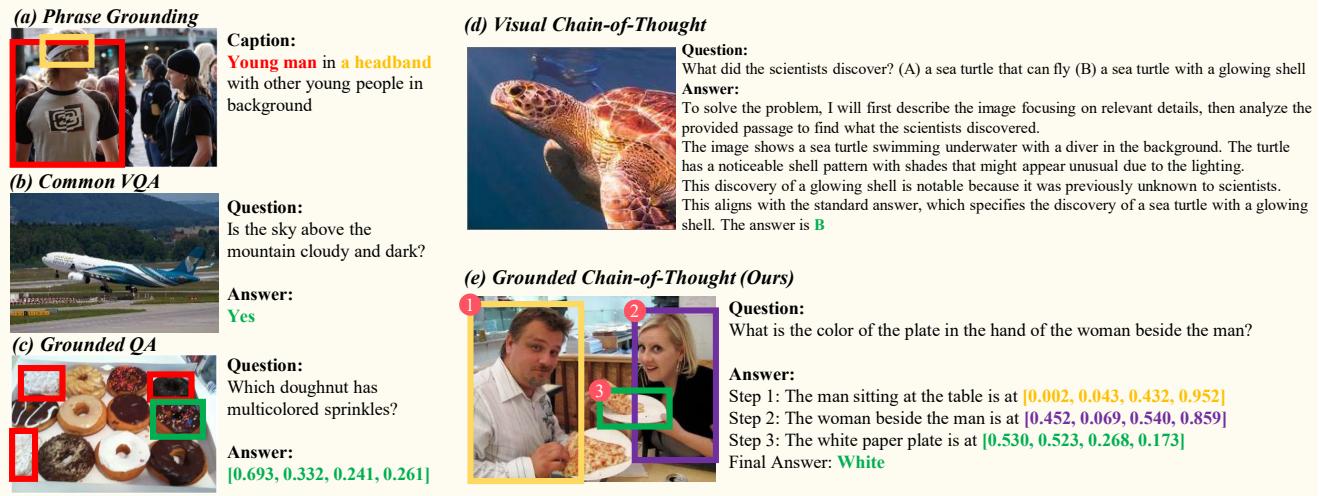


Figure 2. Comparison between relevant vision-language tasks (a-d) and our GCoT (e). Phrase grounding (a) is similar to our GCoT in terms of grounding outputs, but it is only to detect mentioned instances in the caption. Grounded QA (c) also lacks the hidden reasoning steps compared to GCoT. VCoT (d) extends the common VQA of MLLMs (b) via providing more detailed answering thoughts, which is more about knowledge reasoning. In contrast, our GCoT aims to decompose the question into multiple task steps with grounded information, providing intuitive basis for the visual-spatial reasoning of MLLMs.

propose to force the model to ground the related instance in the image when answering, *e.g.*, grounded QA [50], or give more detailed explanations for their answers [31, 45].

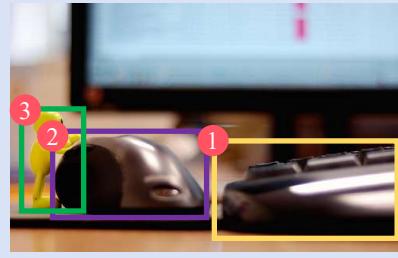
In this paper, we propose to alleviate the visual hallucination of MLLMs via enhancing their visual-spatial reasoning capability, and introduce a new learning task called *grounded chain-of-thought* (GCoT), as shown in Fig.1-(b). Similar to the CoT research for LLMs [18, 39, 43], GCoT is also to help the model make step-by-step reasoning before answering the question. But different from recent visual CoT works [16, 31, 37, 45], which mainly focus on visual knowledge reasoning, GCoT aims to cultivate both visual and spatial perception capabilities of MLLMs, allowing them to ground the key steps for correctly answer the question. This property also make GCoT distinct from some relevant VL tasks, such as grounded QA [50] and visual grounding [4, 17], as shown in Fig.2.

To facilitate GCoT, we also propose a large dataset for the SFT tuning of existing MLLMs, named by *multimodal and grounded chain-of-thought* (MM-GCoT). MM-GCoT consists of 23,028 training examples of three tasks, *i.e.*, *Attribute*, *Judgment*, and *Object*. As shown in Fig.3, an MM-GCoT example has multiple steps with grounding information. In this case, MLLMs can learn to reason key elements in the image for answering, and provide their coordinates as the intuitive basis, yielding a more trustworthy reasoning process. Moreover, GCoT also reserves about 994 examples for the evaluation of the visual-spatial reasoning capability and visual hallucination of MLLMs, which is conducted by three metrics, namely *answer accuracy*, *grounding accuracy* and *answer-grounding consistency*, respectively.

In the experimental section, we first evaluate a set of advanced MLLMs on our MM-GCoT benchmark, including LLaVA-1.5[28], LLaVA-OneVision[22], Qwen2.5-VL[41] and InternVL2.5[3]. From these results, we summarize three notable findings about these MLLMs. Firstly, most MLLMs exhibit low answer-grounding consistency, indicating the great difference between their perception and reasoning. Thus, it is hard to ensure that their answers are what they see. Secondly, this consistency is not related to their grounding or reasoning capability, *i.e.*, an MLLM may performs well on a single task of visual grounding or question answering, but is hard to maintain a good consistency on MM-GCoT. Last but not least, although super-large MLLMs has much stronger multimodal QA accuracies on common benchmarks, but perform very poorly on MM-GCoT, showing obvious data over-fitting during testing. For instance, Qwen-VL-72B shows much worse answer-grounding consistency than its 7B version. These findings imply that visual hallucination related to data overfitting is prominent in existing MLLMs, especially the super-large ones. Afterwards, we further examine the benefit of GCoT by training MLLMs with our MM-GCoT data. The result show that MLLMs enhanced by GCoT capabilities can exhibit much better visual-spatial reasoning while reducing the inconsistency between the answer and grounded proofs, which also show the great potential of GCoT in alleviating visual hallucination.

Overall, our contributions are three-fold:

- We propose a new learning task to alleviate visual hallucination of MLLMs from the perspective of visual-spatial reasoning, termed *grounded chain-of-thought* (GCoT).



Question:
What is the color of the desktop figurine next to the black and brown computer mouse near the computer keyboard?

Answer:

Step 1: The computer keyboard is at [0.540, 0.562, 0.460, 0.367]
 Step 2: The black and brown computer mouse near the keyboard is at [0.114, 0.519, 0.398, 0.327]
 Step 3: The desktop figurine next to the mouse is at [0.050, 0.420, 0.138, 0.407]
 Final Answer: **Yellow**

(a) Attribute

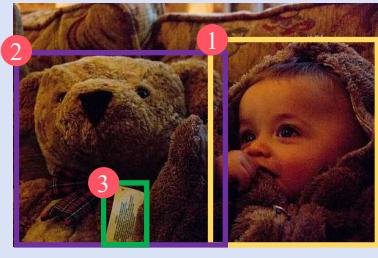


Question:
Is the blue stuffed animal in the child's right hand a bear?

Answer:

Step 1: The child dressed in a costume is at [0.000, 0.096, 0.907, 0.902]
 Step 2: The girl's right hand is at [0.211, 0.336, 0.163, 0.116]
 Step 3: The blue stuffed animal in the girl's right hand is at [0.187, 0.518, 0.336, 0.278]
 Step 4: The blue stuffed animal is a bear.
 Final Answer: **Yes**

(b) Judgement



Question:
What is on the front of the bear next to the baby chewing on a jacket?

Answer:

Step 1: The baby chewing on a jacket is at [0.528, 0.144, 0.472, 0.853]
 Step 2: The bear next to the baby is at [0.000, 0.213, 0.582, 0.784]
 Step 3: The white tag on the front of the bear is at [0.244, 0.740, 0.128, 0.257]
 Final Answer: **Tag**

(c) Object

Figure 3. Examples of the proposed *multi-modal grounded chain-of-thought* (MM-GCoT). MM-GCoT has three splits of examples, namely *Attribute* (a), *Judgement* (b) and *Object* (c). Each example consists of multiple reasoning steps with grounded information, i.e., the spatial coordinates, serving the cultivation of GCoT capability for MLLMs. Meanwhile, MM-GCoT also reserves a set of example for the hallucination evaluation of MLLMs in terms of the metrics of *answer accuracy*, *grounding accuracy* and *answer-grounding consistency*.

- We construct a new dataset called MM-GCoT to facilitate the research of GCoT for MLLMs, which also contains a benchmark to evaluate the hallucination of MLLMs.
- The extensive experiments indicate that notable findings of existing MLLMs in terms of visual hallucination, and also validate the merits of GCoT capability for MLLMs.

2. Related Work

Recent advances in multimodal large language models (MLLMs) [1–3, 22, 28, 41] have demonstrated remarkable capabilities in vision-language tasks, significantly outperforming traditional vision-language models [5, 19, 23]. Despite their impressive performance, MLLMs still suffer from the long-lasting issue of visual hallucination. In traditional vision-language research, visual hallucination primarily denote *language bias* [12, 21, 44, 48], where models tend to predict answers based on question-answer distribution bias rather than the observed visual information. Recent research mainly regards the visual hallucination of MLLMs as the problem of visual shortcoming [7, 30, 36], while the issue of language bias receives less attention. Several benchmark datasets have been developed to evaluate visual hallucination in MLLMs. For instance, POPE [24] introduces object-presence verification tasks. While HallusionBench [9] through professionally designed question pairs and control groups to assess MLLMs’ hallucination

tendencies. However, these benchmarks often rely on simple binary verification metrics, which is hard to judge the reliability of MLLMs’ reasoning process. To this end, we focus more on examining the hallucination of MLLMs in a granular manner, especially about language prior.

In terms of *chain-of-thought* (CoT) research for MLLMs, recent efforts [31, 32, 37, 45] mainly focus on visual explanations or knowledge reasoning. For example, LLaVA-CoT [45] requires an MLLM to output the detailed solution about the problem, and then output the answer. ScienceQA [31] anchors its scientific reasoning in knowledge derived from pre-trained memory, where multimodal contexts activate latent domain information. However, these approaches do not verify whether each reasoning step is correct visual evidence.

Our work is also closely related to visual grounding tasks [4, 17, 20, 33, 35, 50]. Previous visual grounding task like *phrase grounding* aim to locate all objects mentioned in a given caption [17, 33], our GCoT is to analyze the question and form the grounded reasoning steps, and then give the correct answer. *Grounded Question Answering* (GroundedQA) is a extended visual grounding task, which also aims to verify the credibility of MLLM’s outputs by providing the grounding information. However, GroundedQA lacks of the annotated reasoning steps to improve the model’s visual-spatial capability [34, 50]. In contrast, our MM-CoT

Table 1. Statistics of each split of MM-GCoT

Split	Attribute	Judgement	Object	Total
Trainval	5,183	12,849	4,996	23,028
Test	459	206	329	994

Table 2. Comparison between MM-GCoT and VL datasets.

Datasets	TrainVal	Test	CoT	Grounding
Visual7W [50]	229,557	98,382		✓
ScienceQA [31]	16,967	4,241	✓	
MME-CoT [16]	-	1,130	✓	
MM-GCoT	23,028	994	✓	✓

can help the models to identify and ground visual evidence throughout the entire reasoning process. Another recent work similar to ours is VoCoT [25], which also focuses on the grounded reasoning steps for MLLMs. In addition to the difference of the way to construct grounding datasets, our work also differs in the focus of alleviating visual hallucination in MLLMs, and proposes a set of new evaluation metrics for this issue.

The proposed GCoT can be also contributed to the research of *Vision-Language-Action* (VLA) models for embodied applications [47]. By incorporating grounded spatial reasoning into the chain-of-thought process, our approach enables more accurate and interpretable decision-making in complex embodied tasks. This integration is particularly valuable for robotic navigation and manipulation scenarios where precise spatial understanding and step-by-step action planning are essential for successful task completion.

3. Grounded Chain-of-Thought

In this paper, we propose a new learning task for *multimodal large language models* (MLLMs) termed *grounded chain-of-thought* (GCoT). GCoT aims to alleviate the visual hallucination issue of MLLMs via enhancing their grounded visual-spatial reasoning capability. As shown in Fig.2, given a question and an image, GCoT will first let MLLMs to analyze and decompose the task, then reason about each task step and provide the spatial information of task-related elements in the image. Based on these grounded reasoning steps, the MLLM will predict the answer and provide the spatial coordinates as the intuitive basis.

In this context, we can see that GCoT transform the direct answering of MLLMs, originally formulated as a simple mapping function $\mathcal{F} : \mathcal{I}, \mathcal{T} \rightarrow \mathcal{A}$, into a multi-step decision process:

$$P(\mathcal{A}|\mathcal{I}, \mathcal{T}) = \prod_{t=1}^T P(R_t, G_t|\mathcal{I}, \mathcal{T}, \mathbf{G}_{<t}, \mathbf{R}_{<t}), \quad (1)$$

where \mathcal{I} and \mathcal{T} are input image and text. \mathcal{A} represents the final answer. \mathbf{G}_t and \mathbf{R}_t denote the visual evidence and the

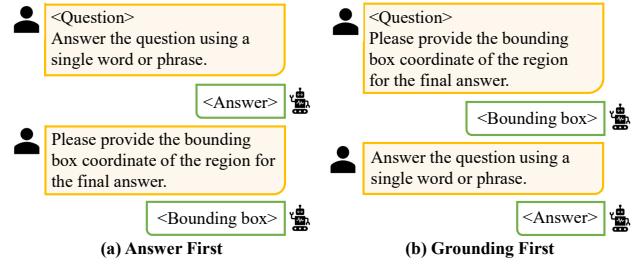


Figure 4. Illustration of two prompting settings for evaluating LLaVA [28] on MM-GCoT. (a) Answer-First: The MLLM generates a textual response, followed by producing the corresponding bounding box in subsequent conversational turns. (b) Grounding-First: The MLLM initially provides a visual bounding box, then responses with a text answer.

textual reasoning state at step t . From Eq.2, we can see the cultivation of GCoT can be accomplished by reinforcement-learning based schemes in addition to the SFT tuning used in this paper, e.g., GRPO [38] used in DeepSeek-R1 [10]. Thus, this setting provides ample room for practitioners to explore RL training on non-GCoT labeled data. Besides, its reasoning step is about the thinking of visual scenes with grounded spatial information, which is also different from existing (visual) CoT tasks. In addition to MLLMs, this setting is also very critical for the exploration of *vision-language-action* (VLA) models in embodied scenarios [47].

4. MM-GCoT: A New Multimodal Dataset

To promote the research of GCoT, we also carefully design and construct a large SFT dataset for MLLMs, termed MM-GCoT. As shown in Tab.2, our MM-GCoT is the only dataset that combines grounding with CoT. Similar to [37, 45], MM-GCoT can be used to train the basic GCoT capability of MLLMs, thereby facilitating the future exploration of RL or weakly-supervised training on more VL data. Besides, MM-GCoT also reserves a part of examples for the visual hallucination of MLLMs.

Dataset Statistics. MM-GCoT contains 23,028 training/validation samples for training GCoT and 994 test samples for the hallucination evaluation of MLLMs. As shown in Fig.1, all samples are categorized into three types, namely *Attribute*, *Judgement* and *Object*. Here, *Attribute* refers to questions providing the object's type and location while inquiring about specific attributes of the target. *Judgement* is to examine the correctness of the given description, while *Object* is the question that focuses on identifying the object category at a specified location.

Dataset Construction. We follow a rigorous four-stage pipeline to build MM-GCoT, which is designed to ensure both reasoning complexity and grounding precision. First, we align region descriptions with object annotations from Visual Genome [20]. To ensure precise spatial correspondence, we employ IoU-based matching between these el-

Table 3. Evaluation results of existing MLLMs on our MM-GCoT benchmark under the *answer-first* prompting setting. “A-Acc” denotes the answer accuracy. “G-Acc” denotes the grounding accuracy. “Consist.” denotes the answer-grounding consistency. The best and second best results are marked in **bold** and underline, respectively.

Method	Attribute			Judgement			Object			Average		
	A-Acc.	G-Acc	Consist.									
LLaVA-7B [28]	68.6	9.2	8.8	83.0	11.2	11.5	58.4	9.1	9.9	70.0	9.8	10.1
LLaVA-13B [28]	69.7	12.0	9.6	83.5	14.6	15.4	58.7	11.6	14.4	70.6	12.7	13.1
LLaVA-OneVision-0.5B [22]	65.6	0.0	0.0	83.0	0.0	0.0	43.5	0.0	0.0	64.0	0.0	0.0
LLaVA-OneVision-7B [22]	69.8	11.2	10.9	89.8	12.1	11.7	65.2	10.5	10.7	<u>74.9</u>	11.3	11.1
LLaVA-OneVision-72B [22]	<u>74.3</u>	13.5	11.0	88.3	17.0	17.9	<u>64.4</u>	19.1	18.5	75.7	16.5	15.8
InternVL2.5-4B [3]	63.4	21.8	18.8	84.5	24.3	20.4	46.2	23.1	20.0	64.7	23.1	19.8
InternVL2.5-8B [3]	61.9	51.9	42.6	<u>89.3</u>	36.4	31.5	48.3	43.2	39.4	66.5	43.8	37.8
InternVL2.5-38B [3]	63.2	53.2	44.3	84.0	<u>45.6</u>	<u>43.5</u>	53.8	42.9	38.9	67.0	47.2	42.2
InternVL2.5-78B [3]	57.9	41.8	33.1	86.9	25.7	26.1	50.8	<u>52.9</u>	47.0	65.2	40.1	35.4
Qwen2.5-VL-3B-Instruct [41]	71.5	<u>57.3</u>	<u>50.0</u>	86.9	39.3	39.8	57.1	<u>48.9</u>	<u>51.7</u>	71.8	<u>48.5</u>	<u>47.2</u>
Qwen2.5-VL-7B-Instruct [41]	73.6	72.5	59.8	87.9	56.3	51.5	57.8	64.1	59.1	73.1	64.3	56.8
Qwen2.5-VL-72B-Instruct[41]	75.5	44.7	41.6	85.9	31.1	32.4	62.9	43.2	43.0	74.8	39.6	39.0

Table 4. Evaluation results of existing MLLMs on our MM-GCoT benchmark under the *grounding-first* prompting setting. “A-Acc” denotes the answer accuracy. “G-Acc” denotes the grounding accuracy. “Consist.” denotes the answer-grounding consistency. The best and second best results are marked in **bold** and underline, respectively.

Method	Attribute			Judgement			Object			Average		
	A-Acc.	G-Acc	Consist.									
LLaVA-7B [28]	59.7	6.3	5.6	82.5	0.5	0.6	35.9	5.5	9.7	59.4	4.1	5.3
LLaVA-13B [28]	<u>69.7</u>	11.5	11.7	<u>83.3</u>	6.1	6.6	<u>51.9</u>	14.6	17.2	<u>68.3</u>	10.8	11.8
LLaVA-OneVision-0.5B [22]	0.0	0.0	0.0	57.3	0.0	0.0	3.0	0.0	0.0	20.1	0.0	0.0
LLaVA-OneVision-7B [22]	69.3	9.2	7.5	79.1	12.1	11.9	38.0	9.1	0.0	62.1	10.1	6.5
LLaVA-OneVision-72B [22]	73.2	17.2	14.0	88.3	13.6	13.5	56.8	17.9	17.7	72.8	16.2	15.1
InternVL2.5-4B [3]	27.2	36.2	22.8	68.0	17.5	18.1	44.6	40.7	14.5	36.6	31.5	18.5
InternVL2.5-8B [3]	54.5	43.1	40.7	76.9	31.9	26.1	37.1	46.2	31.1	56.2	40.4	32.6
InternVL2.5-38B [3]	58.2	71.2	51.1	82.0	57.3	<u>49.5</u>	44.1	51.1	44.2	61.4	59.8	48.3
InternVL2.5-78B [3]	63.4	41.8	33.1	79.1	42.7	41.0	34.0	58.4	<u>37.6</u>	58.9	47.6	37.2
Qwen2.5-VL-3B-Instruct [41]	26.1	<u>82.6</u>	25.1	37.4	66.0	26.8	23.4	55.6	31.3	29.0	68.1	27.7
Qwen2.5-VL-7B-Instruct [41]	48.8	<u>82.6</u>	<u>45.7</u>	80.6	<u>72.8</u>	62.4	26.7	<u>62.3</u>	32.6	52.0	<u>72.6</u>	<u>46.9</u>
Qwen2.5-VL-72B-Instruct[41]	5.2	85.0	5.1	23.3	73.8	23.5	0.3	67.5	0.5	9.6	75.4	9.7

ements. When an object has only one region whose IoU meets the threshold, we consider them to be a match. Secondly, we use the matched objects as nodes to construct a spatial relationship graph based on spatial and semantic relations. To generate multi-step reasoning chains, we iteratively sample the relationship paths from this graph. Thirdly, we use structured templates to gather information of bounding box coordinates, object attributes, and contextual relationships. Afterwards, we apply an LLM [26] to translate the templates into fluent natural language questions. We implement a quality assurance process for both training and test samples. For training data, we use LLM-based verification to check answer accuracy and grounding consistency. Test samples are further validated manually to ensure the highest quality standards.

Evaluation Metrics. In terms of MM-GCoT evaluation, we mainly consider the evaluation for visual hallucination of MLLMs, since existing MLLMs are hard to inspire GCoT outputs without specific tuning. In this case, we focus on measuring the consistency of final answers only. Following previous VQA and Visual Grounding settings [8, 17], we evaluate model performance using three key metrics: *answer accuracy (A-Acc)*, *grounding accuracy (G-Acc)*, and

answer-grounding consistency (Consist.). A-Acc employs text matching between predicted and ground-truth answers, while G-Acc use Acc@0.5 as the metric [17], requiring IoU > 0.5 between predicted and ground-truth boxes. The answer-grounding consistency metric is defined by

$$Con. = \frac{|S_{ca,cb}|}{|S_{ca,cb}| + |S_{ca,wb}| + |S_{wa,cb}|}, \quad (2)$$

where $S_{ca,cb}$, $S_{ca,wb}$ and $S_{wa,cb}$ denote sample sets with correct answer with correct box, correct answer with wrong box, and wrong answer with correct box, respectively.

Evaluation Prompt Settings for Existing MLLMs. Most existing MLLMs are capable of visual grounding, but they are still hard to directly output answer with grounding information [22, 41]. In this case, we design two prompting strategy to inspire their grounded QA capability. As illustrated in Fig.4, the first solution is the **answer-first prompting**, which require the model to provide the answer first and then give the answer coordinates. The second one is **grounding-first prompting**. It lets MLLMs first ground the potential answer, based on which answer the question. In terms of two settings, the answer-first one will be relatively easier, since the MLLM is to detect the answer instance. However,

this setting can check whether the MLLM really see the answer cue in the image. The other grounding-first setting is more challenging but also close to the target of our hallucination evaluation. It can examine whether the MLLM find the key visual element and answer the question based on it. All examined MLLMs will report the three metrics mentioned above on these two settings.

5. Experiments

5.1. Implement Details

We conduct a comprehensive evaluation of leading MLLMs through our GCoT dataset, selecting both pioneering architectures and state-of-the-art performers. The evaluated models include LLaVA (7B, 13B) [28] and LLaVA-OneVision (0.5B, 7B, 72B) [22], along with top-performing Qwen2.5-VL (3B, 7B, 72B) [41] and InternVL2.5 (4B, 8B, 38B, 78B). To enable GCoT adaptation, we replace the original instruction-tuning samples with an equivalent number of examples from our GCoT dataset. Then we train LLaVA GCoT from the scratch using this modified training set following the default settings of LLaVA.

5.2. Quantitative Analysis

Evaluation on MM-GCoT dataset As shown in Tab.3 and Tab.4, we evaluate the performance of 12 state-of-the-art MLLMs on our MM-GCoT dataset using both answer-first and grounding-first strategies. The evaluated models span diverse architectures and parameter scales, including LLaVA [29], LLaVA-OneVision [22], Qwen2.5-VL [41], and InternVL2.5 [3]. We can first observe that existing MLLMs exhibit severe visual hallucination issues. For instance, while LLaVA-OneVision-72B achieves a 75.7% average accuracy, it only maintains 11.1% answer-grounding consistency under the answer-first prompt setting. Notably, even the most advanced MLLM, *e.g.* InternVL2.5-78B, still demonstrates limited accuracy-grounding consistency, achieving only 35.4% on average. Secondly, we observe that visual hallucination does not directly correlate with model scale. The Qwen series shows marginal improvements in answer accuracy with only 1.3% from 3B to 7B, and a mere 1.9% when scaling to 72B. Furthermore, regarding answer-grounding consistency, the Qwen2.5-VL-7B model exceeds its 72B by 18.2%, reinforcing the observation that larger models do not guarantee better visual reasoning reliability. Overall, these experiment results confirm the visual hallucination problem in current MLLMs.

Comparison between Different Subsets. We then analyze the performance across different subsets, *i.e.*, “Attribute”, “Judgement” and “Object”, as shown in Tab.3 and Tab.4. From these tables, we can first observe that the “Attribute” subset demonstrates superior grounding accuracy. For instance, Qwen2.5-VL-7B achieves the ground-

ing accuracy of 82.6% under the grounding-first prompt setting. For “Attribute” task, questions typically contain explicit spatial relationship descriptions, closely aligning with the grounding objectives used in MLLM training, thus yielding optimal grounding accuracy. We can also observe that MLLMs achieve the highest answer accuracy in the “Judgement” task. Specifically, the average answer accuracy achieves 86.1% on average under the answer-first prompt setting. “Judgement” task, which primarily involve fact verification, present relatively straightforward challenges for MLLMs, resulting in higher answer accuracy. Notably, some MLLMs, *e.g.* Qwen2.5-VL-3B, unable to strictly follow the instruction output under the grounding-first prompt setting. It leads to the answer accuracy lower than 50% on the “Judgement” task. Another observation is that MLLMs show the strongest answer-grounding consistency in the “Object” task. For example, the base model InternVL2.5-8B achieves 39.4% answer-grounding consistency in the “Object” task. In “Object” task, once MLLMs successfully identify the target, they can use the corresponding visual region to generate consistent responses, leading to superior consistency metrics. Overall, these varied performance patterns across different task types demonstrate the multi-faceted nature of MLLM capabilities, highlighting how our GCoT dataset effectively evaluates different aspects of multimodal reasoning.

Comparison between Different Prompting Strategies.

We further compare the experiment results between Tab.3 and Tab.4. In the grounding-first setting, models are required to first identify relevant visual regions before answering questions. We can first observe that MLLMs generally achieve higher accuracy in the answer-first approach. For example, LLaVA-7B shows a 8.9% higher accuracy under answer-first compared to the grounding-first prompt setting. This discrepancy is particularly pronounced in smaller-scale models. Notably, Qwen2.5-VL-3B exhibits a substantial accuracy gap of 45.4% between the two approaches. Furthermore, we can observe that while models show improved grounding performance in the grounding-first setting, their answer accuracy deteriorates. For instance, Qwen2.5-VL-7B achieves a 7.9% increase in Acc@0.5 for grounding compared to the answer-first approach, yet suffers an average accuracy drop of 21.2% in answer generation. These findings suggest that correctly question answering may not rely on explicit visual evidence. Overall, these results indicate that MLLM does not consistently use visual and textual information during inference.

Comparison between Different Model Scalings. We then compare the performance of the models from the same series but have different parameters. We can observe that a model’s visual grounding capability does not translate into related reasoning in responses. For instance, while Qwen2.5-VL-72B-Instruct demonstrates superior ground-

Table 5. Comparison between the original MLLMs and LLaVA GCoT. “AF” and “GF” refer to answer-first and grounding-first prompting settings, respectively. “A-Acc” denotes the answer accuracy. “G-Acc” denotes the grounding accuracy. “Consist.” denotes the answer-grounding consistency. The best and second best results are marked in **bold** and underline, respectively.

Method	Prompt Setting	Attribute			Judgement			Object			Average		
		A-Acc.	G-Acc	Consist.									
Qwen2.5-VL-7B-Instruct [41]	AF	<u>73.6</u>	<u>72.5</u>	59.8	87.9	56.3	51.5	57.8	64.1	59.1	73.1	64.3	56.8
Qwen2.5-VL-7B-Instruct [41]	GF	48.8	82.6	45.7	80.6	72.8	<u>62.4</u>	26.7	<u>62.3</u>	32.6	52.0	72.6	46.9
LLaVA-7B [28]	AF	68.6	9.2	8.8	83.0	11.2	11.5	58.4	9.1	9.9	70.0	9.8	10.1
LLaVA-7B [28]	GF	59.7	6.3	5.6	82.5	0.5	0.6	35.9	5.5	9.7	59.4	4.1	5.3
LLaVA-7B GCoT	GCoT	72.8	66.7	56.1	<u>88.3</u>	<u>61.7</u>	56.9	<u>62.3</u>	61.7	61.3	<u>74.5</u>	63.3	<u>58.1</u>
LLaVA-13B [28]	AF	69.7	12.0	9.6	83.5	14.6	15.4	58.7	11.6	14.4	70.6	12.7	13.1
LLaVA-13B [28]	GF	69.7	11.5	11.7	83.3	6.1	6.6	51.9	14.6	17.2	68.3	10.8	11.8
LLaVA-13B GCoT	GCoT	74.7	67.8	<u>58.0</u>	90.3	72.8	68.0	64.1	60.8	<u>59.3</u>	76.4	<u>67.1</u>	61.8

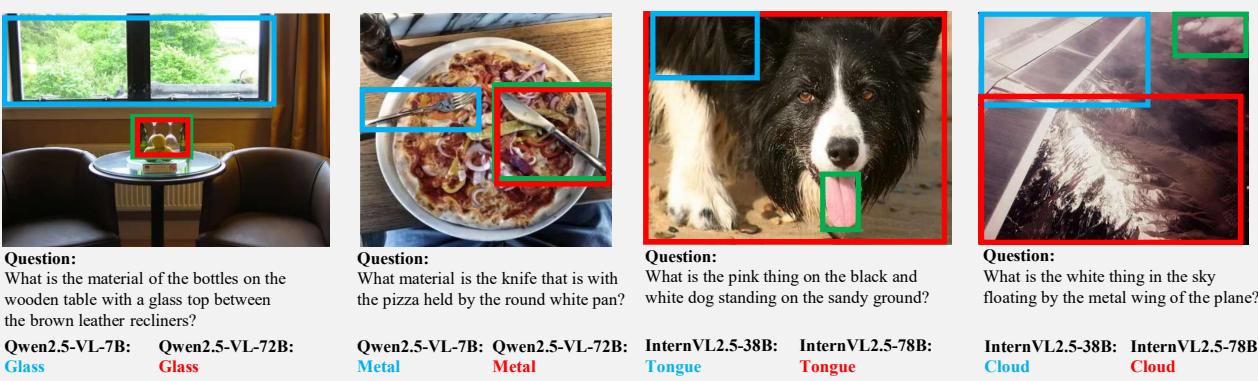


Figure 5. Visualization of the results from Qwen2.5-VL and InternVL2.5 models of different parameter scales under *answer-first* and *grounding-first* promptings, respectively. The **blue**, **red** and **green** bounding boxes represent predictions from base-scale models (Qwen2.5-VL-7B and InternVL2.5-38B), large-scale models (Qwen2.5-VL-72B and InternVL2.5-78B), and ground truth, respectively. These visualizations show that super-large MLLMs exhibit poorer consistency in contrast.

ing performance compared to Qwen2.5-VL-7B-Instruct, achieving a 2.4% improvement on the “Attribute” task under the grounding-first prompt setting, the smaller model surprisingly exhibits better consistency scores with a 40.6% margin. Conversely, the MLLMs’ reasoning ability does not automatically translate into enhanced grounding capability through textual information. For example, while InternVL2.5-38B improves the answer accuracy by 5.5% compared to its 8B counterpart, its grounding performance actually decreases by 0.3% in the “Object” task. Overall, these experimental results prove a gap between visual evidence and textual reasoning in current MLLMs, suggesting the need for more integrated training approaches.

The effects of GCoT training We compare the performance of our LLaVA GCoT with state-of-the-art MLLMs in Tab.5. We can first observe that our GCoT training scheme significantly enhances the consistency and overall performance of MLLMs. Specifically, LLaVA-7B GCoT achieves substantial improvements with a 4.5% increase in accuracy and a remarkable 55.7% boost in consistency compared to the original LLaVA-7B. Furthermore, LLaVA-13B GCoT demonstrates competitive performance against Qwen2.5-VL-7B-Instruct, achieving comparable or better

results across multiple metrics. We can also observe that the improvement in consistency primarily stems from enhanced visual understanding rather than mere grounding capabilities. For example, LLaVA-13B GCoT maintaining similar grounding performance (a slight increase of 2.2%) while dramatically improving consistency by 61.2%. Overall, these results prove that GCoT effectively alleviate the limitation of current MLLMs in using visual information through explicit multimodal alignment.

5.3. Qualitative Analysis

Comparison between Model Series. As shown in Fig.5, we visualize the results of the Qwen2.5-VL and InternVL2.5 models that perform under answer-first and grounding-first prompt setting, respectively. We can first observe that Qwen2.5-VL-7B correctly predicts both the answer and corresponding regions in these two samples. However, while Qwen2.5-VL-72B, correctly provides the answer, it identifies regions that share similar attributes with the target but are not actually relevant to the question. This phenomenon demonstrates that merely scaling up model parameters and enhancing overall capabilities cannot eliminate the hallucinations that exist in MLLMs. Under the

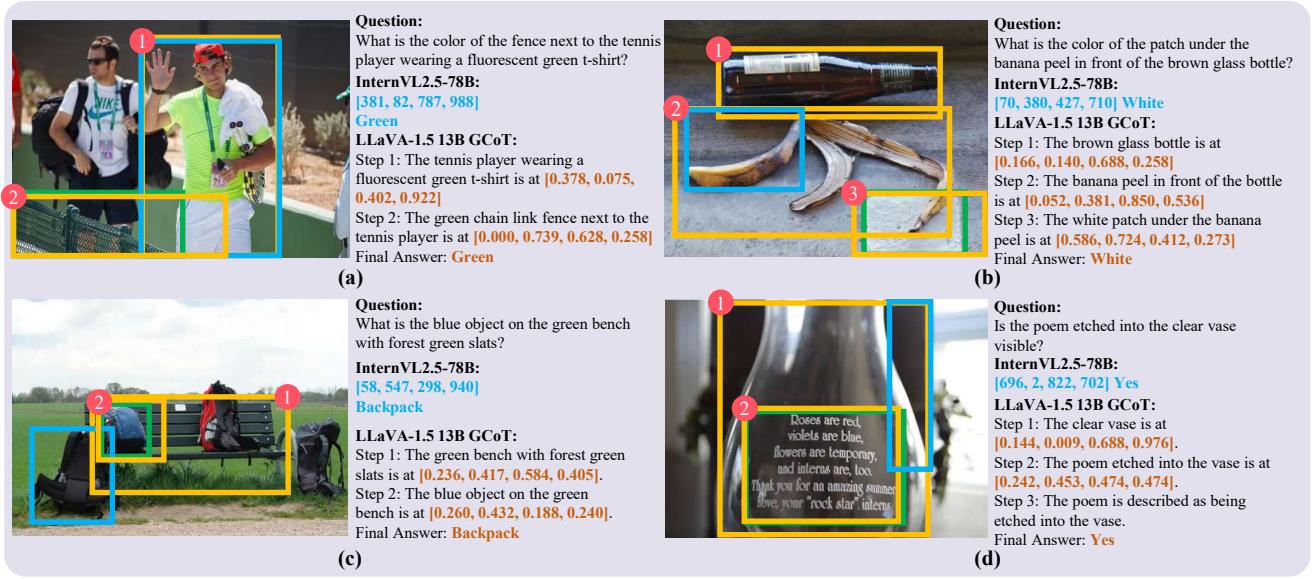


Figure 6. Comparison between LLaVA-1.5 13B GCoT and InternVL2.5-78B. InternVL2.5-78B’s responses and spatial predictions are visualized with blue bounding boxes. For LLaVA-1.5 13B GCoT, we display its answers, chain-of-thought reasoning process, and corresponding regions in yellow. The ground truth regions are indicated by green bounding boxes.

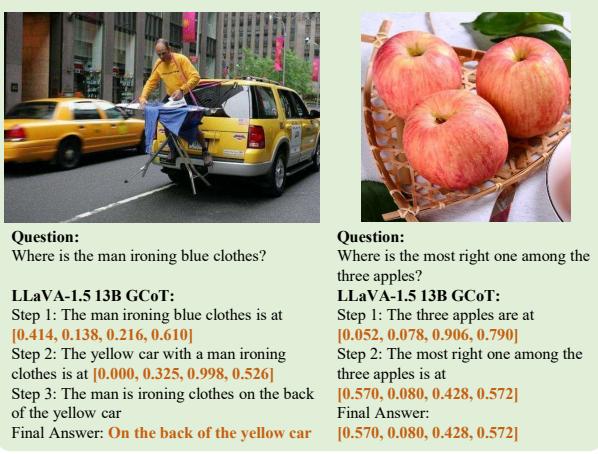


Figure 7. Examples of LLaVA-1.5 13B GCoT on other tasks. After training on our MM-GCoT, LLaVA-1.5 13B can transfer this capability to the tasks like common VQA and Visual Grounding.

grounding-first prompt setting, InternVL2.5 provides correct answers despite identifying mismatched regions. The explicitly given error regions did not cause the model to answer the question incorrectly. This suggests that even larger MLLMs do not necessarily rely on accurate visual evidence to generate their answers. Overall, these experiment results reveal a gap between model scaling and visual grounding capabilities, highlighting the importance of explicit grounding mechanisms for reliable reasoning.

Visualization of LLaVA GCoT. As shown in Fig.6, we compare the performance of LLaVA-1.5 13B GCoT and InternVL2.5-78B through their responses and answer-

grounding consistency. We can first observe that LLaVA-1.5 13B GCoT effectively reduces hallucinations commonly observed in MLLMs. For instance, in Fig.6-(a) and (b), InternVL2.5-78B misidentifies objects with similar attributes and generates responses based on these incorrectly grounded regions. In contrast, LLaVA-1.5 13B GCoT conducts step-by-step reasoning based on precise visual evidence, leading to accurate answers. Furthermore, we observe that the inconsistency between answers and grounding becomes more pronounced in questions that lack direct object descriptions. As demonstrated in Fig.6-(c), even the state-of-the-art InternVL2.5-78B produces confused bounding box predictions when encountering similar objects. However, LLaVA-1.5 13B GCoT successfully disambiguates between multiple potential targets through its chain-of-thought reasoning. In some cases, even powerful MLLMs can exhibit severe hallucinations. For example, in Fig.6-(d), InternVL2.5-78B generates answers based on completely irrelevant regions. In contrast, LLaVA-1.5 13B GCoT accurately identifies the target region while using only 16% of the parameters.

Beyond the comparison, we visualize the reasoning and grounding results in the open-world QA and grounding tasks in Fig.7. From the figure, we can observe that LLaVA-1.5 13B GCoT generalizes well to tasks not included in our GCoT training datasets. Specifically, it can effectively handle open-world questions while providing detailed visual evidence to support its reasoning. Moreover, for referring expression comprehension tasks, the model generates answers with transparent and convincing reasoning processes. Overall, these experimental results comprehen-

sively demonstrate the effectiveness of GCoT in mitigating hallucinations in MLLMs while maintaining strong generalization capabilities.

6. Conclusion

In this paper, we propose a new learning task for MLLMs termed *grounded chain-of-thought* (GCoT). GCoT aims to let MLLMs learn discrete visual reasoning steps with grounded information, thereby alleviating the visual hallucination issue. We also propose a new dataset for MLLMs called MM-GCoT, which can be used to train GCoT capability as well as evaluate the visual hallucination of existing experiments. Extensive experiments are conducted, and the experimental results not only yield notable findings about the hallucination of existing MLLMs, but also show the merits of GCoT towards trustworthy visual reasoning.

References

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *Computing Research Repository (CoRR)*, 2023. 1, 3
- [2] Zhe Chen, Jiannan Wu, Wenhui Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *arXiv preprint arXiv:2312.14238*, 2023.
- [3] 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*, 2024. 1, 2, 3, 5, 6
- [4] Pelin Dogan, Leonid Sigal, and Markus Gross. Neural sequential phrase grounding (seqground). In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4175–4184, 2019. 1, 2, 3
- [5] Zi-Yi Dou, Yichong Xu, Zhe Gan, Jianfeng Wang, Shuohang Wang, Lijuan Wang, Chenguang Zhu, Pengchuan Zhang, Lu Yuan, Nanyun Peng, et al. An empirical study of training end-to-end vision-and-language transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18166–18176, 2022. 3
- [6] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Computing Research Repository (CoRR)*, 2023. 1
- [7] Yuhan Fu, Ruobing Xie, Xingwu Sun, Zhanhui Kang, and Xirong Li. Mitigating hallucination in multimodal large language model via hallucination-targeted direct preference optimization. *arXiv preprint arXiv:2411.10436*, 2024. 3
- [8] Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 6904–6913, 2017. 1, 5
- [9] Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, Dinesh Manocha, and Tianyi Zhou. Hallusionbench: An advanced diagnostic suite for entangled language hallucination & visual illusion in large vision-language models, 2023. 1, 3
- [10] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 4
- [11] Shashank Gupta, Vaishnavi Shrivastava, Ameet Deshpande, Ashwin Kalyan, Peter Clark, Ashish Sabharwal, and Tushar Khot. Bias runs deep: Implicit reasoning biases in person-assigned llms. *arXiv preprint arXiv:2311.04892*, 2023. 1
- [12] Tianyang Han, Qing Lian, Rui Pan, Renjie Pi, Jipeng Zhang, Shizhe Diao, Yong Lin, and Tong Zhang. The instinctive bias: Spurious images lead to illusion in mllms. *arXiv preprint arXiv:2402.03757*, 2024. 3
- [13] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*, 2019. 1
- [14] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*, 2024. 1
- [15] Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046, 2024. 1
- [16] Dongzhi Jiang, Renrui Zhang, Ziyu Guo, Yanwei Li, Yu Qi, Xinyan Chen, Liuhui Wang, Jianhan Jin, Claire Guo, Shen Yan, et al. Mme-cot: Benchmarking chain-of-thought in large multimodal models for reasoning quality, robustness, and efficiency. *arXiv preprint arXiv:2502.09621*, 2025. 2, 4
- [17] Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara Berg. ReferItGame: Referring to objects in photographs of natural scenes. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 787–798, Doha, Qatar, 2014. Association for Computational Linguistics. 2, 3, 5
- [18] Seungone Kim, Se June Joo, Doyoung Kim, Joel Jang, Seonghyeon Ye, Jamin Shin, and Minjoon Seo. The cot collection: Improving zero-shot and few-shot learning of language models via chain-of-thought fine-tuning. *arXiv preprint arXiv:2305.14045*, 2023. 2
- [19] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning (ICML)*, pages 5583–5594, 2021. 3
- [20] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome:

- Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 3, 4
- [21] Baiqi Li, Zhiqiu Lin, Wenzuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. Naturalbench: Evaluating vision-language models on natural adversarial samples. *arXiv preprint arXiv:2410.14669*, 2024. 3
- [22] 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, 2024. 1, 2, 3, 5, 6
- [23] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*, 2023. 3
- [24] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *Computing Research Repository (CoRR)*, 2023. 1, 3
- [25] Zejun Li, Ruiyu Luo, Jiwen Zhang, Minghui Qiu, and Zhongyu Wei. Vocot: Unleashing visually grounded multi-step reasoning in large multi-modal models. *arXiv preprint arXiv:2405.16919*, 2024. 4
- [26] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024. 5
- [27] 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*, 2023. 1
- [28] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1, 2, 3, 4, 5, 6, 7
- [29] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916, 2023. 6
- [30] Shi Liu, Kecheng Zheng, and Wei Chen. Paying more attention to image: A training-free method for alleviating hallucination in lilm. In *European Conference on Computer Vision*, pages 125–140. Springer, 2024. 3
- [31] Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 2022. 1, 2, 3, 4
- [32] Pan Lu, Hritik Bansal, Tony Xia, Jiacheng Liu, Chunyuan Li, Hannaneh Hajishirzi, Hao Cheng, Kai-Wei Chang, Michel Galley, and Jianfeng Gao. Mathvista: Evaluating mathematical reasoning of foundation models in visual contexts. *arXiv preprint arXiv:2310.02255*, 2023. 3
- [33] Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 11–20, 2016. 3
- [34] Dong Huk Park, Lisa Anne Hendricks, Zeynep Akata, Anna Rohrbach, Bernt Schiele, Trevor Darrell, and Marcus Rohrbach. Multimodal explanations: Justifying decisions and pointing to the evidence. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 8779–8788, 2018. 3
- [35] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 1, 3
- [36] Xiaoye Qu, Jiashuo Sun, Wei Wei, and Yu Cheng. Look, compare, decide: Alleviating hallucination in large vision-language models via multi-view multi-path reasoning. *arXiv preprint arXiv:2408.17150*, 2024. 3
- [37] Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Advances in Neural Information Processing Systems*, 37:8612–8642, 2024. 2, 3, 4
- [38] Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Y Wu, et al. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*, 2024. 4
- [39] Qingyi Si, Tong Wang, Zheng Lin, Xu Zhang, Yanan Cao, and Weiping Wang. An empirical study of instruction-tuning large language models in chinese, 2023. 2
- [40] Yinan Sun, Zicheng Zhang, Haoning Wu, Xiaohong Liu, Weisi Lin, Guangtao Zhai, and Xiongkuo Min. Explore the hallucination on low-level perception for mllms. *arXiv preprint arXiv:2409.09748*, 2024. 1
- [41] Qwen Team. Qwen2.5-vl, 2025. 1, 2, 3, 5, 6, 7
- [42] Yining Wang, Mi Zhang, Junjie Sun, Chenyue Wang, Min Yang, Hui Xue, Jialing Tao, Ranjie Duan, and Jieci Liu. Mirage in the eyes: Hallucination attack on multi-modal large language models with only attention sink. *arXiv preprint arXiv:2501.15269*, 2025. 1
- [43] Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2021. 2
- [44] Zongyu Wu, Yuwei Niu, Hongcheng Gao, Minhua Lin, Zhiwei Zhang, Zhifang Zhang, Qi Shi, Yilong Wang, Sike Fu, Junjie Xu, et al. Lanp: Rethinking the impact of language priors in large vision-language models. *arXiv preprint arXiv:2502.12359*, 2025. 3
- [45] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024. 2, 3, 4
- [46] Wenda Xu, Guanglei Zhu, Xuandong Zhao, Liangming Pan, Lei Li, and William Yang Wang. Pride and prejudice: Llm amplifies self-bias in self-refinement. *arXiv preprint arXiv:2402.11436*, 2024. 1

- [47] 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. *arXiv preprint arXiv:2412.14171*, 2024. [4](#)
- [48] Xuesong Zhang, Jia Li, Yunbo Xu, Zhenzhen Hu, and Richang Hong. Seeing is believing? enhancing vision-language navigation using visual perturbations. *arXiv preprint arXiv:2409.05552*, 2024. [3](#)
- [49] Hanzhang Zhou, Zijian Feng, Zixiao Zhu, Junlang Qian, and Kezhi Mao. Unibias: Unveiling and mitigating llm bias through internal attention and ffn manipulation. *arXiv preprint arXiv:2405.20612*, 2024. [1](#)
- [50] Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. Visual7w: Grounded question answering in images. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4995–5004, 2016. [1](#), [2](#), [3](#), [4](#)
- [51] Xianwei Zhuang, Zhihong Zhu, Zhampeng Chen, Yuxin Xie, Liming Liang, and Yuexian Zou. Game on tree: Visual hallucination mitigation via coarse-to-fine view tree and game theory. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17984–18003, 2024. [1](#)
- [52] Xianwei Zhuang, Zhihong Zhu, Yuxin Xie, Liming Liang, and Yuexian Zou. Vasparsc: Towards efficient visual hallucination mitigation for large vision-language model via visual-aware sparsification. *arXiv preprint arXiv:2501.06553*, 2025. [1](#)