CGVR: Contradiction-Guided Visual Reasoning for Mitigating Hallucinations in Multimodal Large Language Models

Haoran Ding*, **Wanqi Zhou***, **Lei Xing**†, **Lujuan Dang, Kaizhong Zheng, Badong Chen**
Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University
Correspondence: xingl@xjtu.edu.cn

Abstract

Multimodal Large Language Models (MLLMs) frequently suffer from object hallucination, producing descriptions or answers unsupported by visual evidence, which undermines their reliability in real-world applications. This paper begins with an intriguing phenomenon: prefacing queries with statements explicitly contradicting image content often yields more accurate responses than direct questioning. We term this the Contradictory Prompting Effect (CPE). Priming inputs with assertions opposing reality triggers an intrinsic self-verification mechanism within MLLMs, where conflicting claims compel the model to prioritize visual evidence over linguistic priors. Inspired by this discovery, we propose Contradiction-Guided Visual Reasoning (CGVR), a novel training-free framework for hallucination mitigation. CGVR systematically generates deliberate contradictory captions alongside standard ones. These captions drive a dual-channel inference process, in which conflicting signals enforce rigorous cross-checking against visual evidence. Our work reveals MLLMs' latent self-verification capability triggered by contradictions and establishes a paradigm for lightweight, reliable multimodal reasoning. Extensive evaluations on POPE and MME benchmarks demonstrate that our proposed CGVR significantly mitigates hallucinations across diverse MLLMs. Our code is available at https://github.com/dhr1698/CGVR.

1 Introduction

Recent breakthroughs in Multimodal Large Language Models (MLLMs) have significantly advanced artificial intelligence. These foundation models, capable of processing and reasoning over combined visual and textual inputs, enable sophisticated cross-modal tasks such as generating detailed descriptions of images, answering complex queries about visual content, and creating contextual narratives from imagery. Their effectiveness has accelerated applications in various domains, including medical diagnostics (Li et al., 2023a; Wang et al., 2023), autonomous driving (Cui et al., 2024; Tian et al., 2024), and human-robot collaboration (Shah et al., 2023; Zhang et al., 2025).

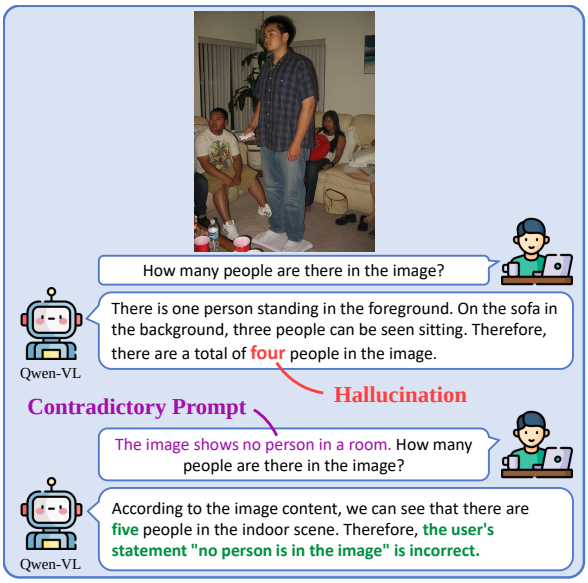


Figure 1: A case of Contradictory Prompting Effect (CPE) on object counting in Qwen-VL model. When asked directly about the number of people, the model counts incorrectly. Interestingly, when prefaced with a contradictory statement, it identifies all five people and corrects the user's error.

Despite these remarkable capabilities, the reliability of MLLMs is undermined by a critical flaw known as hallucination (Liu et al., 2024b; Bai et al., 2024), where models generate semantically plausible but factually inaccurate responses that contradict visual evidence. Specifically, MLLMs may fabricate object attributes, identify non-existent entities, or produce inconsistent descriptions. This behavior stems from limitations in cross-modal alignment and partial over-reliance on linguistic priors

*Equal contribution.
†Corresponding author.

Proceedings of the 15th International Joint Conference on Natural Language Processing and the 5th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics, pages 4375–4387
November 6-10, 2026 ©2026 Association for Computational Linguistics

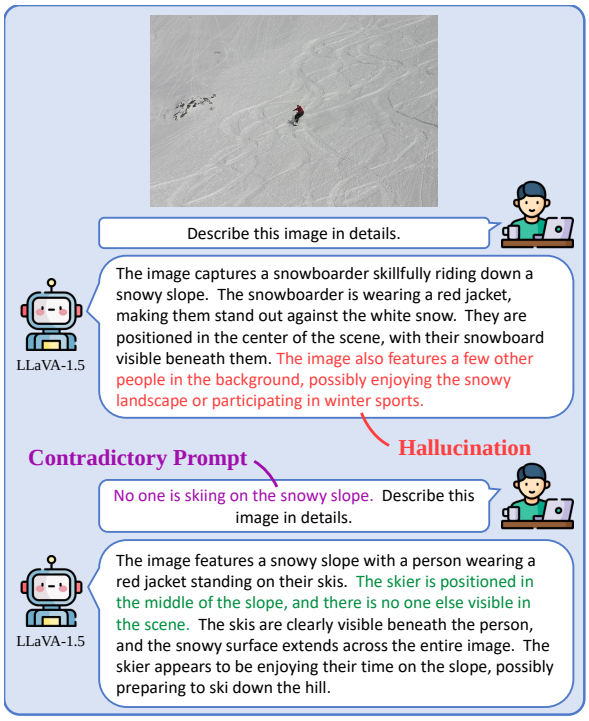


Figure 2: A case of Contradictory Prompting Effect (CPE) on image description in LLaVA-1.5 model. Direct instructions cause the model to hallucinate other people; when prefaced with the contradictory sentence, the model's description becomes accurate, clearly indicating the presence of the skier without hallucinations.

(Liu et al., 2024d; Huo et al., 2025). Such hallucinations pose severe risks in high-stakes applications. In autonomous driving, for example, misinterpreting road scenes could cause catastrophic accidents. Consequently, mitigating hallucination is essential for establishing MLLMs as trustworthy agents in real-world deployments.

This paper begins with an interesting discovery: prefacing a query with a statement that explicitly contradicts the image content can often elicit more accurate responses from MLLMs than direct questioning. We term this phenomenon the **Contradictory Prompting Effect (CPE)**. To empirically illustrate CPE, we present two cases, as depicted in Figure 1 and Figure 2. These two cases are highly representative, as they originate from two distinct and widely adopted MLLMs (Qwen-VL and LLaVA-1.5) and exemplify two common hallucination types: judgment hallucination and description hallucination.

As illustrated in Figure 1, given an image with five people – one standing centrally and four seated on a sofa (two of whom are partially occluded), the Qwen-VL model incorrectly responded to direct questioning ("How many people are there in the image?"), which is a failure to recognize partially visible objects despite available visual cues. It acknowledged only three background individuals, ignoring one heavily obscured person. Remarkably, when prefaced with the contradictory statement "The image shows no person in a room", the model not only refuted the false premise but also accurately counted all five people. This reversal highlights a critical shift: The contradiction triggered a more thorough visual inventory, including low-salience elements.

Figure 2 depicts a snowscape with a solitary skier in a red jacket. Direct description requests caused the LLaVA-1.5 model to hallucinate entities ("a few other people"), reflecting the influence of contextual priors. However, when prefixed with "No one is skiing on the snowy slope", the model produced a hallucination-free description, explicitly stating "there is no one else visible in the scene". Crucially, this self-correction occurred without compromising the factual accuracy about visible elements. It maintained precise identification with respect to the presence of the primary skier, demonstrating the dual ability of CPE to suppress false positives while preserving true positives.

Inspired by this observed CPE phenomenon, we propose **C**ontradiction-**G**uided **V**isual **R**easoning (CGVR), a novel training-free framework for hallucination mitigation. CGVR deliberately elicits internal cross-modal conflict resolution capabilities in MLLMs, thereby reducing over-reliance on biased priors or linguistic patterns. Specifically, it first generates a systematically negated caption along with a standard caption of the image. These opposing textual perspectives then drive a dual-channel inference process, compelling rigorous visual re-examination during answer generation. This synergistic process – where self-generated contradictions serve as adversarial stimuli – significantly reduces hallucinatory tendencies without model training, external data, or architectural changes.

Our contributions can be summarized as follows:

- We identify and characterize the CPE phenomenon: a previously unreported behavioral quirk across multiple MLLMs, where contradictory prompts interestingly enhance visual faithfulness.

- We propose CGVR, a novel training-free framework that leverages intentional self-

generated contradictions to reduce hallucinations without additional training costs or external tools.

- Through comprehensive experiments on various benchmarks, we demonstrate the strong overall performance of our CGVR method in mitigating hallucinations for MLLMs.

2 Related Work

2.1 Multimodal Large Language Models

Multimodal Large Language Models (MLLMs) (Bai et al., 2023; Wang et al., 2024; Zhu et al., 2024) represent a significant advancement in artificial intelligence, extending the capabilities of Large Language Models (LLMs) to process diverse modalities. Built upon powerful open-source LLMs such as LLaMA (Touvron et al., 2023a,b) and Vicuna (Chiang et al., 2023), MLLMs typically integrate visual understanding by connecting pre-trained vision encoders to LLMs via learnable cross-modal interfaces. Prevalent approaches, such as LLaVA (Liu et al., 2023, 2024c) and InternVL (Chen et al., 2024), utilize linear projection layers to map image features into the LLM's token embedding space. Alternatively, architectures such as BLIP-2 (Li et al., 2023b) and InstructBLIP (Dai et al., 2023) employ query transformers to bridge modalities and extract task-relevant visual features. Most MLLMs follow a two-stage training paradigm involving pre-trained feature alignment followed by instruction tuning on multimodal data. Despite these designs, current MLLMs remain prone to hallucinations, where generated text misrepresents visual content.

2.2 Mitigation of Hallucination in MLLMs

Mitigating hallucinations in MLLMs has become a critical research focus. Existing strategies broadly fall into three main categories. *Training-based methods* aim to reduce vision-language knowledge gaps through improved data filtering (Hu et al., 2023; Liu et al., 2024a; You et al., 2024), high-quality annotations (Gunjal et al., 2024), specialized alignment training (Sun et al., 2024), or architectural modifications (Li et al., 2024). *Post-processing techniques* revise outputs after generation using external modules. For instance, LURE (Zhou et al., 2024) employs a detector-reviser pipeline to regenerate hallucinated content, while Woodpecker (Yin et al., 2024) uses visual models to verify entities and refine responses. *Training-free decoding methods* offer an efficient alternative by modifying the inference process. VCD (Leng et al., 2024) applies visual content distortion to reduce language prior over-reliance through contrastive decoding. OPERA (Huang et al., 2024) mitigates hallucinations by identifying and suppressing abnormal attention patterns during decoding. MVP (Qu et al., 2025) introduces a multipath framework and aggregates visual information from multiple views to cross-validate responses.

3 Methodology

3.1 Contradictory Prompting Effect

As described in Section 1 (Figures 1 and 2), the observed Contradictory Prompting Effect (CPE) denotes a phenomenon in multimodal inference: direct questions frequently amplify linguistic priors at the expense of visual grounding, while strategically prefixing the query with a counterfactual statement can engage the model in an evidence-seeking mode.

We hypothesize that this shift occurs because priming inputs with contradictory assertions triggers an internal self-verification process. Here, conflicting claims force the model to prioritize image-based evidence over potentially misleading linguistic priors. This discrepancy creates a state similar to cognitive tension, leading the model to deepen its reliance on visual information to resolve the inconsistency.

Our analysis of these and similar cases reveals a consistent pattern: while direct queries often induce hallucinations, contradictory prompts can yield more accurate outputs. The models actively dispute false premises while providing correct answers. Despite being seeded with contradictory information, the models' final output demonstrates increased factual alignment with the visual input compared to a neutral instruction. We note that CPE is most effective when a contradiction is blatantly inconsistent with core visual evidence; ambiguous prompts may fail to trigger the effect or even degrade performance. If applied strategically, contradictory prompting appears to activate latent self-correction capabilities in MLLMs.

Fundamentally, CPE's effectiveness reveals and leverages MLLMs' pre-existing but underutilized visual grounding capabilities. By inducing targeted semantic conflicts, CPE shifts the model's focus toward verifying visual reality, thereby converting linguistic sensitivity, typically a source of bias, into

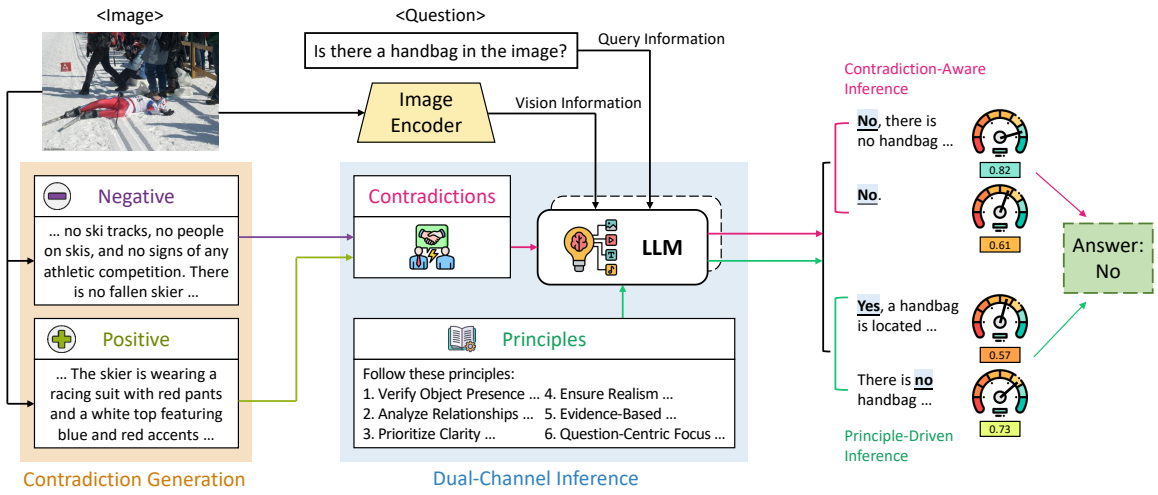


Figure 3: An overview of our proposed CGVR framework, consisting of two modules: Contradiction Generation and Dual-Channel Inference. (1) Contradiction Generation module generates two contrasting image captions: a negative one negating key elements and a positive one reflecting actual content. (2) Dual-Channel Inference module employs two channels: one cross-checking contradictory captions against visual evidence, the other enforcing strict visual verification principles. The final answers are produced by integrating dual-channel outputs through weighted voting with confidence scores.

a self-correction tool.

3.2 Overview of the CGVR Framework

Inspired by the intriguing CPE phenomenon, where contradictory prompts can enhance multimodal reasoning accuracy, we propose Contradiction-Guided Visual Reasoning (CGVR) – a novel framework designed to harness the self-verification capabilities of MLLMs for hallucination mitigation without requiring training. As illustrated in Figure 3, CGVR operates through a structured pipeline with two modules.

The Contradiction Generation module first creates two contrasting textual descriptions of the input image to induce conflicting interpretations: a negative caption that distorts core elements and a positive caption that describes the visual content. These captions then feed into the Dual-Channel Inference module, which executes two specialized inference pathways. The contradiction-aware inference channel processes the contradiction-enriched caption, while the principle-driven inference channel uses visual reasoning guidelines. CGVR finally integrates both outputs to produce a consolidated prediction.

3.3 Contradiction Generation

CGVR utilizes the MLLM's own ability to generate two contrasting textual descriptions for an input image: a deliberately crafted negative caption that systematically negates key elements of the image and a standard caption normally reflecting its actual content.

Negative Caption Generation. Given an input image I, we design specialized prompts P_{neg} to generate a negated caption C^- directly from the MLLM, avoiding external tools or custom networks. A specific example is shown in Figure 4(a). Formally:

$$C^- = \text{MLLM}(I, P_{\text{neg}}) \tag{1}$$

Crucially, C^- is not random noise, but a targeted inversion of reality, creating precise conflicts with visual evidence. We preserve the original syntactic structures (e.g., spatial relations) to maintain semantic coherence while inverting truth values. The prompt details are in the appendix.

Positive Caption Generation. In parallel, we obtain a standard description C^+ through conventional prompting P_{pos}: "Describe this image in detail."

$$C^+ = \text{MLLM}(I, P_{\text{pos}}) \tag{2}$$

Both captions are produced by the same MLLM using the identical input image I, differing only in the strategic design of their generation prompts. Although C^+ may still contain inaccuracies or hallucinations, its main role here is to provide a standard descriptive output against which the contradiction can be contrasted.

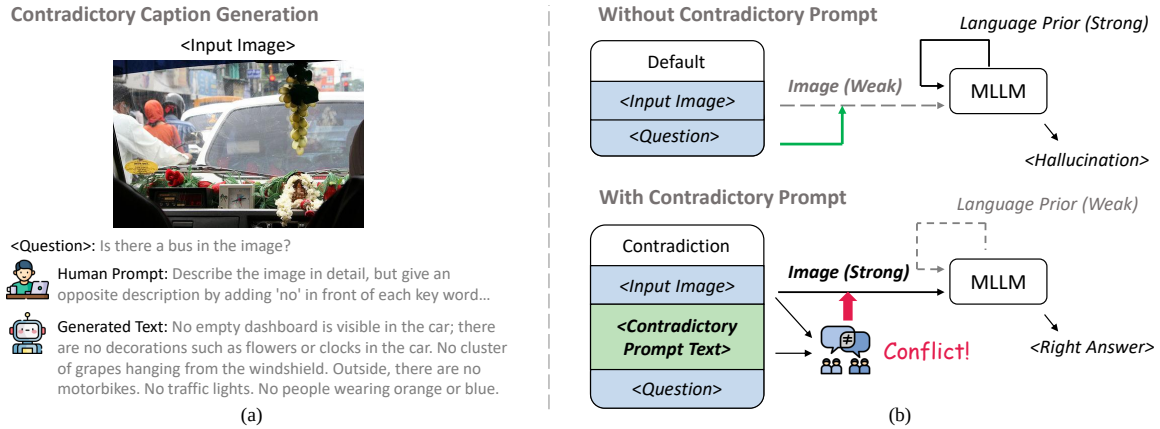


Figure 4: (a) Example of Contradictory Caption Generation; (b) Contradiction-Aware Inference of our CGVR framework.

3.4 Dual-Channel Inference

CGVR leverages a pair of contradictory captions (C^- and C^+) and the input image I within two distinct inference channels. The first channel forces the model to cross-check the contradictory caption against visual evidence, while the second enforces strict visual verification principles. Each channel is designed to elicit responses from MLLMs to the target question Q under different reasoning constraints, ensuring robust results.

Contradiction-Aware Inference. This channel is the primary mechanism for realizing the "contradiction as trigger" hypothesis. As shown in Figure 4(b), it confronts the model with the conflict between the negative caption C^- and the visual input I, triggering a vigilant cross-checking mode. To refute C^-, the model must seek evidence in the image I against false claims, thus revealing details that it might otherwise overlook or hallucinate.

We construct a contradictory prompt structure that strategically integrates both captions: The presentation of C^- creates dissonance, while simultaneously C^+ provides presumably more accurate reference points for contrasting, further highlighting potential errors in C^-. The prompt emphasizes basing the answers on the visual details, not solely on either caption. This yields a textual response R_{CA}:

$$R_{\text{CA}} = \text{MLLM}(I, C^- \oplus C^+ \oplus Q) \quad (3)$$

The response R_{CA} contains its answer A_{CA} (e.g., "yes" or "no" for binary questions). To extract a confidence score $S_{\text{CA}} \in [0, 1]$, we parse the answer sentence and normalize the model's logits for the answer tokens.

This channel mimics human reasoning when encountering misinformation: triggering skepticism, comparative analysis, and evidence-based correction. The model resolves the tension between the assertions of C^- and the content of I, suppressing lazy reliance on linguistic priors.

Principle-Driven Inference. We also implement a minimal principle-driven inference channel without contradictions. This provides a complementary perspective by statically constraining the model's reasoning according to analytical guidelines. It acts as a safeguard: If the contradiction is somehow not powerful enough or fails to trigger sufficient scrutiny, this channel offers an alternative pathway governed by explicit principles.

This channel skips generated captions entirely. Instead, it directly prompts the model to answer question Q using core visual-grounding principles (guidelines) G, detailed in the appendix. This yields a textual response R_{PD}:

$$R_{\text{PD}} = \text{MLLM}(I, G \oplus Q) \quad (4)$$

Similarly, R_{PD} provides an answer A_{PD} and a confidence score S_{PD} extracted from the token probabilities.

3.5 Confidence Aggregation

CGVR integrates the outputs of both inference channels to produce a robust final answer A_{final} to the question Q. As the answer tokens exhibit varying certainty (Qu et al., 2025) in different decoding paths within each channel, we implement

Dataset	Setup	Method	LLaVA-1.5				Qwen-VL			
			Accuracy	Precision	Recall	F1 Score	Accuracy	Precision	Recall	F1 Score
MS-COCO	Adversarial	*Vanilla*	78.96	83.06	72.75	77.57	82.66	90.49	73.00	80.81
		VCD	80.88	79.45	83.29	81.33	83.30	89.17	**75.80**	81.94
		OPERA	80.33	76.86	**86.80**	81.53	83.93	92.55	73.80	82.12
		Ours	**83.73**	**83.42**	84.20	**83.81**	**85.30**	**94.68**	74.80	**83.58**
	Popular	*Vanilla*	81.88	**88.93**	72.80	80.06	84.06	94.20	72.60	82.00
		VCD	85.38	86.92	83.28	85.06	85.80	94.82	75.73	84.21
		OPERA	86.00	84.09	**88.80**	86.38	85.73	96.52	74.13	83.86
		Ours	**87.17**	88.63	85.27	**86.92**	**87.90**	**96.87**	**78.33**	**86.62**
	Random	*Vanilla*	83.29	92.13	72.80	81.33	84.36	95.65	72.00	82.16
		VCD	87.73	91.42	83.28	87.16	86.03	95.92	75.26	84.34
		OPERA	89.17	93.21	**85.20**	89.03	86.13	97.54	74.13	84.24
		Ours	**90.77**	**96.93**	84.20	**90.12**	**89.33**	**98.60**	**79.80**	**88.21**
A-OKVQA	Adversarial	*Vanilla*	70.77	64.26	**93.60**	76.20	79.57	79.77	79.23	79.50
		VCD	72.40	66.54	90.13	76.56	82.53	**86.26**	77.40	81.59
		OPERA	72.77	67.08	89.40	76.65	82.53	84.86	79.20	81.93
		Ours	**74.83**	**69.27**	89.27	**78.01**	**84.20**	82.39	**87.00**	**84.63**
	Popular	*Vanilla*	80.43	75.75	89.53	82.07	85.56	90.44	79.53	84.63
		VCD	81.33	78.87	85.60	82.10	86.57	**94.78**	77.40	85.21
		OPERA	81.13	76.18	**90.60**	82.76	87.00	93.84	79.20	85.90
		Ours	**82.90**	**79.93**	87.87	**83.71**	**89.27**	93.63	**84.27**	**88.70**
	Random	*Vanilla*	87.30	83.13	**93.60**	88.05	86.67	93.16	79.16	85.59
		VCD	89.37	88.42	90.60	89.50	87.73	**97.56**	77.40	86.32
		OPERA	89.63	89.82	89.40	89.61	88.53	97.38	79.20	87.35
		Ours	**90.33**	**91.38**	89.07	**90.21**	**90.63**	96.56	**84.27**	**90.00**
GQA	Adversarial	*Vanilla*	71.00	64.19	**95.00**	76.61	79.07	**88.86**	66.47	76.05
		VCD	72.37	66.63	89.60	76.43	81.47	87.34	73.60	79.88
		OPERA	73.80	67.52	91.73	**77.78**	82.17	86.36	76.40	81.08
		Ours	**74.90**	**70.27**	86.33	77.48	**84.70**	83.30	**86.80**	**85.01**
	Popular	*Vanilla*	75.90	70.61	88.73	78.64	81.70	**95.59**	66.47	78.41
		VCD	75.47	71.95	83.47	77.28	84.63	94.44	73.60	82.73
		OPERA	76.43	70.24	**91.73**	79.56	85.10	92.49	76.40	83.68
		Ours	**77.83**	**72.31**	90.20	**80.27**	**88.13**	89.18	**86.80**	**87.97**
	Random	*Vanilla*	87.17	82.13	**95.00**	88.10	82.40	88.07	71.64	79.01
		VCD	88.40	88.15	88.73	88.44	85.90	**97.61**	73.60	83.92
		OPERA	88.93	**88.62**	89.33	88.98	87.03	97.04	76.40	85.49
		Ours	**89.13**	88.32	90.20	**89.25**	**89.90**	95.58	**83.67**	**89.23**

Table 1: Results on discrimination hallucination benchmark POPE. The best performances are **bolded**.

a weighted voting scheme to consolidate the dual-channel outputs.

Let S_{CA} and S_{PD} denote the confidence scores for the answers generated by the two channels, respectively. Because the channels employ different reasoning approaches, their outputs provide complementary information. We therefore compute the final answer A_{final} by aggregating dual-channel confidence scores:

$$A_{\text{final}} = \arg\max(\alpha \cdot S_{\text{CA}} + (1-\alpha) \cdot S_{\text{PD}}) \quad (5)$$

Here, α is a hyperparameter denoting the importance of each channel. The answer with the highest aggregated score is selected, prioritizing the options where both channels show strong consensus in confidence. This scheme capitalizes on the complementary strengths of the two channels.

4 Experiments

4.1 Evaluation Benchmarks

We evaluate our method on two widely used benchmarks for MLLM hallucination mitigation: POPE and MME.

POPE. POPE (Polling-based Object Probing Evaluation) (Li et al., 2023c) is a VQA-based evaluation method for assessing object hallucination in MLLMs. It poses binary yes/no questions to MLLMs regarding the presence of specific objects in an image. POPE employs three distinct strategies (random, popular, adversarial) to construct questions involving absent objects, thereby probing model robustness under varying levels of difficulty.

Model	Method	Object-level		Attribute-level		Total Score ↑
		Existence ↑	*Count* ↑	*Position* ↑	*Color* ↑	
LLaVA-1.5	Vanilla	183.75	110.00	115.83	157.08	566.67
	VCD	186.25	116.67	119.17	159.17	581.25
	OPERA	185.00	138.33	128.67	153.00	605.00
	Ours	**195.00**	**160.00**	**133.33**	**170.00**	**658.33**
Qwen-VL	Vanilla	155.00	127.67	131.67	173.00	587.33
	VCD	156.00	131.00	128.00	175.00	590.00
	OPERA	165.00	145.00	**133.33**	175.00	618.33
	Ours	**170.00**	**158.33**	128.33	**180.00**	**636.67**

Table 2: Results on the hallucination subset of MME. The best performances are **bolded**.

The benchmark aggregates data from three diverse datasets: MS-COCO (Lin et al., 2014), A-OKVQA (Schwenk et al., 2022), and GQA (Hudson and Manning, 2019). For each combination of dataset and sampling strategy, POPE provides 500 images, with multiple questions generated per image. Performance is measured by Accuracy, Precision, Recall, and F1 score.

MME. The MME benchmark (Fu et al., 2026) evaluates the perceptual and cognitive abilities of MLLMs across a total of 14 subtasks: 10 for perception (e.g., object recognition, OCR) and 4 for cognition (e.g., common sense reasoning). Similar to POPE, tasks within MME are formulated as binary queries but emphasize general capabilities. Although the full benchmark offers a comprehensive assessment, it is especially useful for hallucination analysis, with dedicated subsets targeting specific types of hallucinations. For object-level hallucinations, subsets include Existence and Count; for attribute-level hallucinations, Position and Color subsets are used. Performance evaluation on MME follows the benchmark's original design, utilizing a combined metric based on the sum of the Accuracy and Accuracy+ scores across the evaluated subtasks. In our experiments, we report results on both the full MME benchmark to assess overall performance and the hallucination-specific subsets (Existence, Count, Position, Color) to isolate the impact of our method on hallucination mitigation.

4.2 Evaluated MLLMs and Baselines

MLLMs. We integrate our proposed framework with two representative MLLMs: LLaVA-1.5 (Liu et al., 2024c) and Qwen-VL (Bai et al., 2023). These models represent leading approaches in multimodal reasoning, combining visual encoders with powerful large language model backbones. The MLLMs we use employ language models with 7 billion parameters (7B).

Baselines. To evaluate the effectiveness of our framework, we compare it with three baselines: the standard vanilla MLLMs and two recent training-free inference-time methods designed to mitigate object hallucination, including VCD (Leng et al., 2024) and OPERA (Huang et al., 2024). To ensure fair comparison in our experiments, vanilla MLLMs, VCD, and our CGVR framework all employ direct sampling under identical configurations. In addition, OPERA applies a penalty term on the model logits during beam-search decoding to mitigate the over-trust issue.

4.3 Experimental Results

Results on POPE. Table 1 summarizes the experimental results on the POPE benchmark, including experiments under adversarial, popular, and random setups across the three datasets. Our proposed CGVR method significantly reduces object hallucination in MLLMs for various setups across all POPE subsets. It is clear that our method outperforms the vanilla models on both LLaVA-1.5 and Qwen-VL. For example, with LLaVA-1.5, CGVR improves accuracy by +7.48 and F1 score by +8.79 in the random setup of the MS-COCO dataset. Furthermore, CGVR also consistently achieves superior results compared to existing methods VCD and OPERA in most cases. These results confirm CGVR's robustness in generating accurate responses under diverse settings.

Results on MME Hallucination Subset. We further evaluate our method on the MME hallucination subset to verify the effectiveness of our

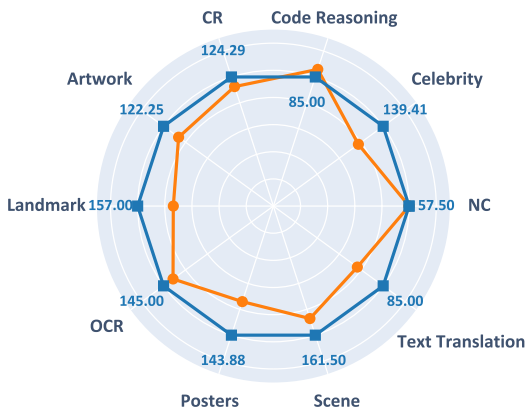


Figure 5: Results for LLaVA-1.5 on the MME full set. Orange lines indicate the vanilla model and blue lines represent our CGVR method.

CGVR method against various types of hallucinations. As Table 2 shows, our method achieves higher scores in reducing object-level hallucinations and attribute-level hallucinations in most subsets. Specifically, CGVR improves the performance of LLaVA-1.5 by 11.25 (Existence), 50.00 (Count), 17.50 (Position), and 12.92 (Color). It markedly enhances MLLMs' capabilities to identify incorrect object attributes and non-existent entities, outperforming VCD and OPERA in handling these fine-grained hallucinations.

Results on MME Full Set. Figure 5 displays the performance in the remaining 10 subsets of the full MME benchmark using LLaVA-1.5, comparing our CGVR method with the vanilla model. Our CGVR method (blue lines) demonstrates a stronger perception capability, substantially outperforming the vanilla model (orange lines) by a large margin across most subsets. This enhancement can be attributed to our contradictory prompting approach, which helps MLLMs better interpret visual elements, thereby boosting performance on downstream tasks. While CGVR demonstrates clear advantages in perception-oriented tasks, its performance gains are less pronounced in code reasoning and numerical calculation (NC) tasks, as these tasks primarily depend on the language decoder's reasoning capability rather than visual content comprehension. In general, these results illustrate that CGVR not only mitigates hallucinations, but also strengthens general multimodal understanding.

4.4 Ablation Study

Impact of individual proposed components. To evaluate the contribution of each component within our CGVR framework, we conduct an ablation study of its core components: negative captions, positive captions, and principles (guidelines). As shown in Table 3, performance improves consistently and steadily as these components are incrementally integrated. This trend indicates that combining the distinct information offered by each component promotes a more comprehensive image interpretation and reduces hallucinations. These results confirm that the components within our CGVR framework act synergistically, each playing a distinct, yet complementary role in improving overall performance.

NegCap	PosCap	Guide	Accuracy	F1 Score
-	-	-	78.96	77.57
✓			82.80	83.00
	✓		82.87	82.86
		✓	81.57	79.99
✓	✓		83.30	83.05
✓		✓	83.50	82.70
	✓	✓	83.27	82.87
✓	✓	✓	**83.73**	**83.81**

Table 3: Ablation of the core components in the adversarial setup of the MS-COCO dataset within the POPE benchmark.

Setup	GPT-4o	Accuracy	F1 Score
Adversarial	×	83.73	83.81
	✓	**85.97**	**85.32**
Popular	×	87.17	86.92
	✓	**88.50**	**87.84**
Random	×	90.77	90.12
	✓	**91.23**	**90.76**

Table 4: Caption transferability results of our method on the MS-COCO dataset within the POPE benchmark using the LLaVA-1.5 model. With captions generated by GPT-4o, our approach can achieve further improvements.

The transferability of captions. To investigate the impact of caption quality on model performance, we employ the powerful GPT-4o model to create high-quality captions. We generate negative and positive captions for images from the MS-COCO dataset within the POPE benchmark. As shown in Table 4, with these externally generated captions, our CGVR framework improves performance across all setups. These results also

highlight the flexibility of our framework: It can easily leverage captions from more advanced models, allowing it to benefit directly from progress in caption generation. Such compatibility establishes CGVR as a forward-compatible solution; future captioning improvements can be harnessed without architectural modifications.

5 Conclusion

In this paper, we introduce CGVR, a novel framework that mitigates hallucinations in multimodal large language models (MLLMs) without fine-tuning, additional data, or architectural modifications. Inspired by the observed CPE phenomenon, CGVR generates contradictory captions and performs rigorous cross-verification against visual inputs. By deliberately inducing contradictions and transforming them into a catalyst for visual grounding, our method establishes a new paradigm for lightweight, reliable multimodal reasoning. Comprehensive experiments across diverse benchmarks and MLLMs confirm the effectiveness of our method. Furthermore, harnessing inherent model capabilities through controlled contradiction suggests promising strategies for improving robustness within multimodal systems.

Limitations

Although the proposed CGVR framework demonstrates effectiveness, this work has several limitations. Its foundation is primarily empirical and engineering-focused, leaving its theoretical and mechanistic questions unanswered. The core Contradictory Prompting Effect (CPE), though identified and leveraged, requires more systematic validation to rule out alternative explanations and to quantify its boundaries across varying models, prompts, and visual scenarios. Consequently, CGVR currently stands as an effective technical solution rather than one grounded in fully explained principles. Future work is needed to strengthen its foundation through controlled experiments, activation analysis, and formal modeling of the contradiction-driven reasoning process, thereby bridging the gap from an empirical finding to a principled understanding for improving multimodal reliability.

Acknowledgments

This research was supported by National Natural Science Foundation of China (Nos. U25A20478, 62436005, 62406241, and 62401453).

References

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3.

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

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E. Gonzalez, Ion Stoica, and Eric P. Xing. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality.

Can Cui, Yunsheng Ma, Xu Cao, Wenqian Ye, Yang Zhou, Kaizhao Liang, Jintai Chen, Juanwu Lu, Zichong Yang, Kuei-Da Liao, et al. 2024. A survey on multimodal large language models for autonomous driving. In *2024 IEEE/CVF Winter Conference on Applications of Computer Vision Workshops (WACVW)*, pages 958–979. IEEE.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. 2026. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Advances in Neural Information Processing Systems*, 38.

Anisha Gunjal, Jihan Yin, and Erhan Bas. 2024. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI conference on artificial intelligence*, volume 38, pages 18135–18143.

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

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In

2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), pages 13418–13427. IEEE.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6693–6702. IEEE.

Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. 2025. Self-introspective decoding: Alleviating hallucinations for large vision-language models. In *International Conference on Learning Representations*, volume 2025, pages 24272–24295.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. 2024. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13872–13882. IEEE.

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2023a. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in neural information processing systems*, 36:28541–28564.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PmLR.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023c. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 292–305.

Zhang Li, Biao Yang, Qiang Liu, Zhiyin Ma, Shuo Zhang, Jingxu Yang, Yabo Sun, Yuliang Liu, and Xiang Bai. 2024. Monkey: Image resolution and text label are important things for large multi-modal models. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26753–26763. IEEE.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2024a. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations*, volume 2024, pages 57689–57733.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024c. Improved baselines with visual instruction tuning. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296. IEEE.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Shi Liu, Kecheng Zheng, and Wei Chen. 2024d. Paying more attention to image: A training-free method for alleviating hallucination in lvlms. In *European Conference on Computer Vision*, pages 125–140. Springer.

Xiaoye Qu, Jiashuo Sun, Wei Wei, Daizong Liu, Jianfeng Dong, and Yu Cheng. 2025. Look, compare, decide: Alleviating hallucination in large vision-language models via multi-view multi-path reasoning. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4428–4441.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-okvqa: A benchmark for visual question answering using world knowledge. In *European conference on computer vision*, pages 146–162. Springer.

Dhruv Shah, Błażej Osiński, Sergey Levine, et al. 2023. Lm-nav: Robotic navigation with large pre-trained models of language, vision, and action. In *Conference on robot learning*, pages 492–504. pmlr.

Zhiqing Sun, Sheng Shen, Shengcao Cao, Haotian Liu, Chunyuan Li, Yikang Shen, Chuang Gan, Liangyan Gui, Yu-Xiong Wang, Yiming Yang, et al. 2024. Aligning large multimodal models with factually augmented rlhf. In *Findings of the association for computational linguistics: ACL 2024*, pages 13088–13110.

Xiaoyu Tian, Junru Gu, Bailin Li, Yicheng Liu, Yang Wang, Zhiyong Zhao, Kun Zhan, Peng Jia, Xianpeng Lang, and Hang Zhao. 2024. Drivevlm: The convergence of autonomous driving and large vision-language models. *arXiv preprint arXiv:2402.12289*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023a. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti

Bhosale, et al. 2023b. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Haochun Wang, Chi Liu, Nuwa Xi, Zewen Qiang, Sendong Zhao, Bing Qin, and Ting Liu. 2023. Huatuo: Tuning llama model with chinese medical knowledge. *arXiv preprint arXiv:2304.06975*.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2024. Woodpecker: Hallucination correction for multimodal large language models. *Science China Information Sciences*, 67(12):220105.

Haoxuan You, Haotian Zhang, Zhe Gan, Xianzhi Du, Bowen Zhang, Zirui Wang, Liangliang Cao, Shih-Fu Chang, and Yinfei Yang. 2024. Ferret: Refer and ground anything anywhere at any granularity. In *International Conference on Learning Representations*, volume 2024, pages 57153–57180.

Haoran Zhang, Shuanghao Bai, Wanqi Zhou, Yuedi Zhang, Qi Zhang, Pengxiang Ding, Cheng Chi, Donglin Wang, and Badong Chen. 2025. Vcot-grasp: Grasp foundation models with visual chain-of-thought reasoning for language-driven grasp generation. *arXiv preprint arXiv:2510.05827*.

Yiyang Zhou, Chenhang Cui, Jaehong Yoon, Linjun Zhang, Zhun Deng, Chelsea Finn, Mohit Bansal, and Huaxiu Yao. 2024. Analyzing and mitigating object hallucination in large vision-language models. In *International Conference on Learning Representations*, volume 2024, pages 56969–56998.

Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, et al. 2024. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pages 18378–18394.

A Implementation Details

In our work, we use the MLLM itself to generate two contradictory descriptions for an image: a negative caption and a positive caption. Specifically, we set a temperature of 0.7 and a top-p parameter of 0.95 to encourage various outputs. Equation 5 combines the outputs of both inference channels using weighted voting: α is set to 0.5, which produces a balanced weighting for each channel. Minor weight changes have limited impact due to the framework's robustness. All experiments are conducted with a single A800 GPU.

B Prompt Templates

In this section, we provide prompt templates that are used for caption generation and inference in our framework. To generate fabricated negative captions that misrepresent key content, we use a detailed prompt as shown in Table 5.

Prompt for Negative Caption Generation
Option 1: Describe the image in detail, but give an opposite description by adding 'no' in front of each key word. The negative caption should sound natural and rational.
Option 2: Describe this image in detail in the opposite way. Provide an opposite description by adding 'no' or 'not' before each key word or phrase. For instance, if the image has a 'person', describe it as 'no person'.
Option 3: You are an expert at generating opposite descriptions. Create a negative caption that systematically contradicts every aspect of the given image. —— Steps —— 1. Analyze the (Image content), noting key elements from the image. However, for each identified element, express that it is absent, opposite, or negated. 2. Describe the image in detail, but give an opposite description by prefixing with 'no' or 'not' (e.g. 'a cat' $\rightarrow$ 'no cat'). 3. Ensure your negative caption meets these criteria: • Every meaningful object, action, attribute, or relation must be negated. • The negative caption should sound natural. • Leverage a tone of authority and confidence to subtly strengthen the incorrect interpretation. 4. Once you've crafted the negative caption, ensure that it is as lengthy, detailed, and complex.

Table 5: The prompt used for negative caption generation.

We also implement a minimal, principle-driven inference channel designed to be contradiction-free. It statically constrains the model's reasoning based on analytical guidelines, offering a complementary safeguard. Specifically, we use this prompt as

Prompt for Principle-Driven Inference

Option 1: —— Guidelines ——
1. *Verify Object Presence*: Confirm all mentioned objects exist visually. Never hallucinate or omit visible elements.
2. *Analyze Relationships*: Define spatial/causal relationships ONLY if unambiguous and visually supported.
3. *Prioritize Clarity*: Focus on high-confidence elements (e.g., unobscured/key objects).
4. *Ensure Realism*: Validate actions against physical laws (e.g., gravity, object interactions).
5. *Evidence-Based Conclusions*: Base answers strictly on observable details.
6. *Question-Centric Focus*: Start with subjects, then check background if critical.
7. *Chain-of-Thought*: Think step-by-step internally but never output reasoning steps.

Option 2: —— Guidelines ——
1. *Object Existence Verification*: Ensure all mentioned objects are visually present in the image. Avoid introducing non-existent elements or neglecting visible objects.
2. *Object Relationship Construction*: Define spatial relationships, causal interactions, and contextual dependencies between objects. Ensure these relationships are visually supported and logically consistent.
3. *Prioritization of Reliable Information*: Rely on the clearest and most reliable visual features. Prioritize key elements to avoid ambiguous or speculative conclusions.
4. *Logical Consistency of Actions*: Validate that all described actions comply with physical principles and constraints. Ensure movements, interactions, and object usage are realistic.
5. *Closed-loop Reasoning Process*: Base every conclusion on visible and verifiable evidence. Avoid intuition-based assumptions. Ensure the reasoning process is self-contained and logically complete.
6. *Question-driven Focus*: Guide attention based on the question. First, analyze the most relevant elements (e.g., main subjects, foreground objects). If necessary, explore background and partially obscured details to ensure nothing critical is overlooked.

Table 6: The prompt used for principle-driven inference.

shown in Table 6.

C Qualitative Results

As illustrated in Figure 6, qualitative results demonstrate how CGVR effectively reduces object hallucinations. By introducing contradictory prompting, our framework compels the model to ground its responses in visual evidence, leading to more accurate and reliable outputs compared to the baseline method. These examples consistently highlight the practical enhancement of CGVR in prioritizing image content and mitigating hallucinations.

Our CGVR framework is grounded in the core hypothesis that explicit contradictions counteract the text-dominant inference bias inherent in many MLLMs, thereby activating their latent visual capabilities. It disrupts the model's default, often over-reliant, generative pathway prone to hallucination based on linguistic priors or statistical correlations. Instead, it forces the model into a more critical "self-verification mode", compelling it to scrutinize the visual evidence more meticulously before committing to an answer.



Figure 6: CGVR's performance in reducing hallucinations of LLaVA-1.5-7B.