Yang *et al*. *Intell. Robot.* **2025**, *5*(2), 248-75
**DOI:** 10.20517/ir.2025.13

Intelligence & Robotics

**Research Article**

Open Access

# Retrieve-then-compare mitigates visual hallucination in multi-modal large language models

**Dingchen Yang**[1], **Bowen Cao**[2], **Sanqing Qu**[3], **Fan Lu**[3], **Shangding Gu**[4,5], **Guang Chen**[3,4]

[1]School of Automotive Studies, Tongji University, Shanghai 201804, China.
[2]The Chinese University of Hong Kong, Hong Kong 999077, China.
[3]Department of Computer Science and Technology, Tongji University, Shanghai 201804, China.
[4]School of Computation, Information and Technology, Technische Universität München, München 80333, Germany.
[5]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA 94704, USA.

**Correspondence to:** Prof. Guang Chen, Department of Computer Science and Technology, Tongji University, 4800 Cao'an Road, Jiading District, Shanghai 201804, China. E-mail: guangchen@tongji.edu.cn

**How to cite this article:** Yang, D.; Cao, B.; Qu, S.; Lu, F.; Gu, S.; Chen, G. Retrieve-then-compare mitigates visual hallucination in multi-modal large language models. *Intell. Robot.* **2025**, *5*(2), 248-75. http://dx.doi.org/10.20517/ir.2025.13

**Received:** 15 Dec 2024 **First Decision:** 24 Feb 2025 **Revised:** 8 Mar 2025 **Accepted:** 10 Mar 2025 **Published:** 25 Mar 2025

**Academic Editor:** Alex Khang PH **Copy Editor:** Pei-Yun Wang **Production Editor:** Pei-Yun Wang

## Abstract

Multi-modal large language models (MLLMs) demonstrate remarkable success in a range of vision-language tasks. However, they are prone to visual hallucinations, where their textual responses diverge from the provided image. Inaccurate visual understanding poses risks to the practical applications of MLLMs. Are MLLMs oblivious to accurate visual cues when they hallucinate? Our investigation indicates that the visual branch of MLLMs may advocate both erroneous and accurate content equally, highlighting a high level of uncertainty. To address this issue, we propose retrieval contrastive decoding (RCD), a training-free method that leverages analogous visual hallucinations, which are induced by images sharing common semantic and appearance characteristics, to mitigate visual hallucinations. Specifically, RCD retrieves relevant images to serve as references for MLLMs, and compares their visual content with the test image through confidence score subtraction. Additionally, RCD coordinates the correction of hallucinations from both the visual and textual branches of MLLMs by adaptively scaling the subtracted scores. Experiments on public hallucination benchmarks demonstrate the efficacy of RCD in mitigating visual hallucinations for three state-of-the-art MLLMs, surpassing other advanced decoding strategies. Furthermore, we validate the effectiveness of RCD in enhancing the capability of MLLMs to comprehend complex and potentially hazardous situations in real-world traffic scenarios. RCD enhances the accuracy of MLLMs in understanding real-world scenes and improves their capability for reasoning, thereby enhancing the reliability of MLLMs in real-world applications.

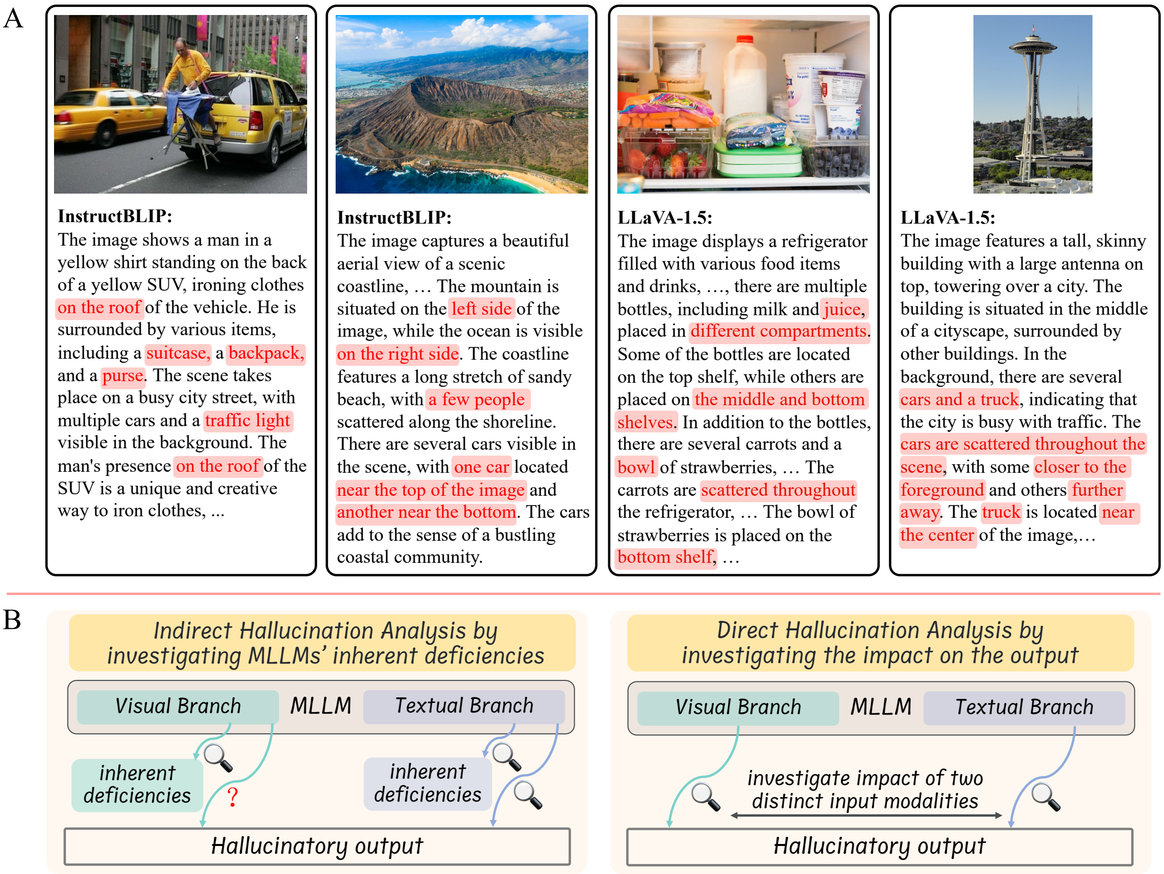
**Keywords:** Vision language model, visual hallucination, autonomous driving

## 1. INTRODUCTION

Multi-modal large language models (MLLMs) have emerged as dominant forces in vision-language tasks[1–8], showcasing remarkable advancements in comprehending a wide array of visual concepts and reasoning with common sense. Pioneering work has also explored using MLLMs as the "brain" of embodied agents[9] and autonomous driving systems[10,11], leveraging their strong visual perception and reasoning capabilities to facilitate high-performance interactions between robotic systems and the real world. Despite their impressive capabilities, state-of-the-art MLLMs are susceptible to visual hallucination[12–19], wherein they inaccurately interpret visual inputs. Specifically, MLLMs can generate conflicting or fabricated content that diverges from the provided image, and may overlook crucial visual details. As illustrated in Figure 1A, leading MLLMs, such as LLaVA-1.5[2] and InstructBLIP[3], often hallucinate non-existent objects (e.g., traffic lights, people, and trucks) and inaccurate locations (e.g., a man on the roof of a car). Inaccurate visual understanding negatively affects subsequent reasoning processes and poses risks for real-world applications of MLLMs. Thus, investigating and mitigating visual hallucinations in MLLMs forms the foundation for enhancing the reliability of intelligent robotic systems in real-world applications.

Understanding the origins of visual hallucinations is paramount for their reduction. Previous studies highlight several flaws in MLLMs, such as insufficiently distinctive visual features[18], the image-text modality gap[14], biased feature aggregation patterns[13,20], and the reliance on superficial language patterns in the training data[15,17]. However, they have not investigated how these flaws specifically lead to the hallucinatory outputs. To address this gap, we develop an end-to-end analytical method to investigate the effects of two distinct input modalities, i.e., images and text, on the output of MLLMs by decoupling the influence of each modality. The distinction between existing studies and our research is depicted in Figure 1B. This investigation suggests a different perspective on visual hallucinations that MLLMs may not be entirely oblivious to accurate visual cues when they produce hallucinations; rather, their predictions reflect an uncertainty between hallucinatory and accurate content. This is evidenced by our observation that the visual branch of MLLMs tends to assign close confidence scores to both accurate and erroneous token candidates, which are referred to as visually deceptive candidates. Assigning considerable positive confidence scores to inappropriate token candidates increases their likelihood of being sampled, thereby leading to visual hallucinations.

The most straightforward approach to distinguish accurate token candidates from hallucinatory ones is directly adjusting the predicted confidence score distribution, prioritizing accurate content over hallucinations. However, this objective lies beyond the capabilities of existing distribution post-processors, particularly the contrastive decoding (CD)-based methods[15,21]. While these methods are effective at reducing the uni-modal bias inherent in the language decoder, we observe that their side effects may exacerbate hallucinations originating from the visual branch of MLLMs, i.e., further promoting visually deceptive candidates. Commencing with the hypothesis that similar images may induce analogous visual hallucinations, we proceed to analyze the shift in confidence score distribution when replacing the test image with retrieved alternatives, and observe moderate changes in the scores of visually deceptive candidates, whereas the scores for accurate candidates exhibit more significant variations. This hypothesis is further supported by quantitative experiments using samples drawn from the VQAv2 validation set[22]. Experimental results demonstrate that LLaVA-Next[23], a state-of-the-art MLLM, exhibits hallucinations in 70% of the test samples. Among the test samples where hallucinations occur, more than two-thirds demonstrate that similar images induce analogous visual hallucinations. Leveraging this phenomenon, we introduce a training-free approach named retrieval CD (RCD). During the inference stage of MLLMs, RCD first retrieves relevant images to serve as visual references for MLLMs. Next, it contrasts the visual cues in the references with those of the test image by subtracting the predicted confidence scores. The subtracted scores are then added to the scores predicted by the test image. This confidence score calibration process suppresses the visually deceptive candidates and promotes the accurate candidates. Additionally, RCD retains the ability to debias erroneous language priors by adaptively scaling the subtracted scores. Thus, RCD is capable of mitigating hallucinations originating from both the visual and textual branches



**InstructBLIP:**
The image shows a man in a yellow shirt standing on the back of a yellow SUV, ironing clothes on the roof of the vehicle. He is surrounded by various items, including a suitcase, a backpack, and a purse. The scene takes place on a busy city street, with multiple cars and a traffic light visible in the background. The man's presence on the roof of the SUV is a unique and creative way to iron clothes, ...

**InstructBLIP:**
The image captures a beautiful aerial view of a scenic coastline, … The mountain is situated on the left side of the image, while the ocean is visible on the right side. The coastline features a long stretch of sandy beach, with a few people scattered along the shoreline. There are several cars visible in the scene, with one car located near the top of the image and another near the bottom. The cars add to the sense of a bustling coastal community.

**LLaVA-1.5:**
The image displays a refrigerator filled with various food items and drinks, …, there are multiple bottles, including milk and juice, placed in different compartments. Some of the bottles are located on the top shelf, while others are placed on the middle and bottom shelves. In addition to the bottles, there are several carrots and a bowl of strawberries, … The carrots are scattered throughout the refrigerator, … The bowl of strawberries is placed on the bottom shelf, …

**LLaVA-1.5:**
The image features a tall, skinny building with a large antenna on top, towering over a city. The building is situated in the middle of a cityscape, surrounded by other buildings. In the background, there are several cars and a truck, indicating that the city is busy with traffic. The cars are scattered throughout the scene, with some closer to the foreground and others further away. The truck is located near the center of the image,…

A
B
Indirect Hallucination Analysis by investigating MLLMs' inherent deficiencies
Visual Branch
MLLM
Textual Branch
inherent deficiencies
?
inherent deficiencies
Hallucinatory output
Direct Hallucination Analysis by investigating the impact on the output
Visual Branch
MLLM
Textual Branch
investigate impact of two distinct input modalities
Hallucinatory output

**Figure 1.** (A) An illustration of visual hallucinations in MLLMs' responses. Leading MLLMs, such as InstructBLIP and LLaVA-1.5, often produce image descriptions that contain inaccuracies, commonly referred to as visual hallucinations. These visual hallucinations can include incorrect object categories (nouns), incorrect spatial relationships (prepositions and adverbs), and inappropriate visual attributes (adjectives). Such errors can diminish the reliability of MLLMs in real-world applications. The present samples are from the LLaVA-bench-in-the-wild dataset, with inaccurate content highlighted in red. (B) Previous research has identified inherent deficiencies in MLLMs, but has not explored how these deficiencies contribute to the hallucinatory responses. While recent studies have explored the impact of the textual branch of MLLMs on hallucinations, the influence of the visual branch has been overlooked. This study addresses this gap by conducting a direct analysis of hallucinations, examining the respective contributions of both the visual and textual branches to hallucinatory responses. MLLMs: Multi-modal large language models.

of MLLMs.

We validate the effectiveness of our proposed method using publicly available hallucination evaluation benchmarks, including MME[24], polling-based object probing evaluation (POPE)[25], WHOOPS[26] and LLaVA-Bench-in-the-wild[27]. RCD significantly improves the performance of LLaVA-Next, LLaVA-1.5, and InstructBLIP on these benchmarks, outperforming existing advanced decoding strategies for hallucination mitigation. To evaluate the capability of RCD in enhancing MLLMs for real-world applications, we develop a comprehensive traffic scenario understanding benchmark. This benchmark features images depicting complex real-world traffic situations, including accidents and significant risk factors, under diverse weather and lighting conditions. We devise six categories of questions to cover a wide range of tasks, including visual recognition, commonsense reasoning, and knowledge-intensive tasks. Our benchmark presents substantial challenges to MLLMs, and RCD enhances the visual recognition and reasoning capabilities of three MLLMs, improving their reliability in practical application.

Our main contributions are three-fold:

- We provide quantitative and qualitative evidence to demonstrate that MLLMs can identify accurate visual cues even amid visual hallucinations. Additionally, we reveal that analogous hallucinations can occur among similar images, and this phenomenon can be exploited to discern accurate visual content. These findings suggest new perspectives for hallucination mitigation.
- We introduce a novel approach to mitigate visual hallucinations in MLLMs. Our approach RCD first retrieves relevant images to serve as references for MLLMs. By comparing the visual cues in the visual references with the test image, RCD discerns accurate visual cues. This process is analogous to the memory mechanism in the human brain, enabling MLLMs to refer to relevant information during the visual recognition process. RCD can be integrated seamlessly into various MLLMs without requiring model retraining.
- Experiments on public hallucination benchmarks demonstrate the superiority of RCD, enhancing the performance of three leading MLLMs and outperforming existing methods. Additionally, we curate a real-world traffic scenario comprehension benchmark, encompassing various challenging tasks. Experiments on our proposed benchmark validate that RCD improves the capabilities of MLLMs in visual recognition and reasoning.

## 2. RELATED WORK

### 2.1. Visual hallucinations and their origins

In traditional computer vision (CV), the image hallucination task typically refers to the process of generating or reconstructing high-quality images [28]. In the context of MLLMs, visual hallucination [12] refers to the issue wherein the descriptive textual content diverges from the visual input. These erroneous responses may exhibit fabrication, contradictions, or scarce specificity to the provided image. Initial investigations primarily address object-level visual hallucinations [17,25,29], focusing solely on inaccurate nouns. This problem is subsequently extended to a finer granularity [18,24,30], including errors in visual attributes, spatial relationships, physical states, activities, numbers, and low-level visual perception tasks [31] regarding degraded images.

The origins of visual hallucination stem from various sources. Some highlight flaws within the visual encoder of MLLMs, such as limited image resolution [23], under-distinctive visual representations that lack visual details [18], and poor cross-modal representation alignment [14]. Others emphasize deficiencies within the language decoder, such as biased attention score distribution [13,20], adherence to superficial syntactical patterns (e.g., frequent answers [19] and contextual co-occurrence of object names [15,17]), the overwhelming parametric knowledge [32], and error snowballing [17,33]. However, the specific mechanisms by which these deficiencies result in hallucinatory outputs (the predicted confidence scores for erroneous token candidates) have not been examined. In this study, we investigate the genesis of visual hallucination by analyzing the direct impact of the visual and textual input on the predicted confidence scores of MLLMs. Experimental results reveal that the visual branch of MLLMs tends to equally advocate both erroneous and accurate token candidates amid visual hallucinations. This observation suggests new perspectives that visual hallucinations can be reduced by straightforwardly adjusting the predicted confidence scores, prioritizing accurate visual content over hallucinations.

### 2.2. Mitigating visual hallucination

#### *2.2.1 Parameter tuning*

Approaches for mitigating visual hallucinations in MLLMs through parameter tuning can be categorized into supervised fine-tuning (SFT) methods and preference optimization techniques. Effective SFT-based methods include curating multi-modal instruction tuning dataset with distracting instructions [19], and the provision of extra supervision to promote image-text feature alignment [14] or the distinctiveness of visual features [7]. On the other hand, preference optimization methods [34–36] construct multi-modal pairwise preference data comprising accurate and erroneous responses and optimize MLLMs using the direct preference optimization (DPO) loss [37]. These methods incorporate inferior responses during training, which may enhance their ability to suppress hallucinations. Nonetheless, for methods relying on SFT or DPO, the training cost becomes prohibitive for large-scale MLLMs. Furthermore, excessive parameter tuning may impair some of the strengths of MLLMs, such as the capability to provide detailed descriptions[19], when the training recipe is suboptimal.

#### *2.2.2 Model ensemble*

Integrating knowledge from other models compensates for MLLMs' own shortcomings. Feasible methods include improving the object recognition accuracy through ensembling object detectors[38], and obtaining distinctive image features by ensembling various pretrained vision encoders[5]. Another line of work utilizes a language model to post-hoc revise visual hallucinations[16,17]. Key challenges within this paradigm include tailoring interfaces for various task-specific models, and automating their selection based on the hallucination categories.

#### *2.2.3 Decoding strategy*

Intervening the decoding process of MLLMs is a more efficient approach compared to model training and ensemble. For instance, OPERA[13] directly discards the candidates that may skew subsequent content toward hallucination and reelects the others. visual contrastive decoding (VCD)[15] and its variants[21,39] extend CD[40,41], which aims to mitigate factual hallucinations in large language models (LLMs), to the vision domain. This line of work distorts the visual input to amplify the language priors, and downgrades the candidates advocated merely by the language priors through logit subtraction. Thus, they are capable of reducing the erroneous language bias inherent in the decoder of MLLMs. However, hallucinations originating from the visual branch of MLLMs have not been examined. As explained in Section 3.2.1, this kind of hallucination may be exacerbated by VCD-based methods. In this study, we investigate the respective impact of the visual and textual input modalities on the hallucinatory content produced by MLLMs. Based on our findings, we propose an approach to mitigate visual hallucinations arising from both the visual and textual branches of MLLMs.

## 3. VISUAL HALLUCINATION ANALYSIS

In this section, we first investigate the genesis and the characteristics of visual hallucinations, as well as the characteristics of the hallucinated content. Commence with a fundamental question: to what extent are MLLMs unaware of accurate visual cues amid hallucinations? We devise an end-to-end experiment pipeline and present our findings in the following sections.

### 3.1. Background and visual hallucination analysis pipeline

Leading MLLMs[2–5,7,23] incorporate auto-regressive language models[42,43], which repeatedly select the next token from their vocabulary $\mathcal{V}$ based on the predicted probability of each token candidate $x_i$,

$$p_\theta(x_i|\boldsymbol{v},\boldsymbol{x},\boldsymbol{y}_{<t}) = \frac{\exp\left(\boldsymbol{h}_t \cdot E_c(x_i)\right)}{\sum_{x'\in\mathcal{V}} \exp\left(\boldsymbol{h}_t \cdot E_c(x')\right)} \in (0,1) \tag{1}$$

where $\boldsymbol{v}$ is the visual input, $\boldsymbol{x}$ and $\boldsymbol{y}_{<t}$ are the prompt and past generated tokens, respectively. $E_c(x_i)$ is the token embedding of candidate $x_i$ in the language model head. $(\cdot)$ is the inner product operator. $\boldsymbol{h}_t$ is the hidden state predicted by the last transformer[44] block of the language model $LLM_\theta$,

$$\boldsymbol{h}_t(\boldsymbol{v},\boldsymbol{x},\boldsymbol{y}_{<t}) = LLM_\theta([VE(\boldsymbol{v});TE(\boldsymbol{x};\boldsymbol{y}_{<t})]) \tag{2}$$

where $VE(\cdot)$ denotes the visual encoder and the cross-modal connector. $TE(\cdot)$ is the input text embedding layer. The confidence score $(\boldsymbol{h}_t \cdot E_c(x_i))$ manifests the significance of $x_i$'s semantics in $\boldsymbol{h}_t$. According to Equation (1), token candidates with higher confidence scores will obtain higher probability, such that they are more likely to be selected.

In this study, we pose the following questions: When visual hallucinations occur, are MLLMs completely ignorant of the accurate visual cues, and is it possible to help them distinguish the accurate content from

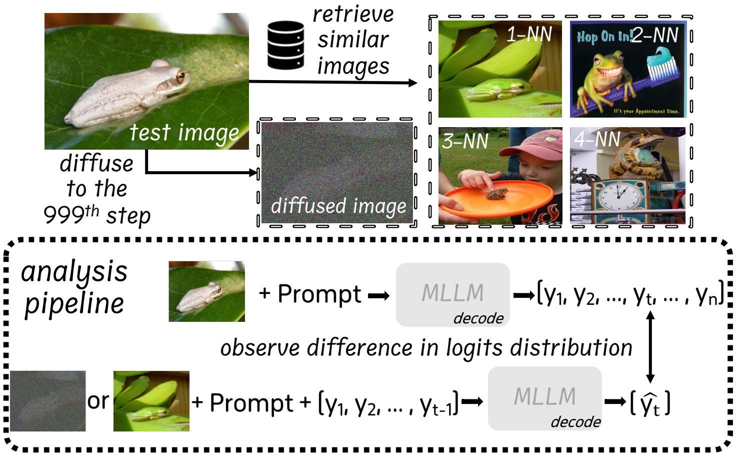

retrieve similar images
test image
diffuse to the 999th step
diffused image
1-NN
2-NN
3-NN
4-NN
analysis pipeline
+ Prompt
MLLM
decode
[y1, y2, ..., yt, ... , yn]
observe difference in logits distribution
or
+ Prompt + [y1, y2, ... , yt-1]
[ŷt]

**Figure 2.** Experiment pipeline for investigating the impact of visual and textual input modality on the hallucinatory output. At each decoding step $t$, the test image $\boldsymbol{v}^\tau$ is replaced with alternative images $\boldsymbol{v}'$ while keeping the textual prefix constant. Next, we assess the difference in output confidence scores (i.e., logits) between $y_t$ and $\hat{y}_t$ to demonstrate the impact of the visual input. This test image is taken from the OpenImages validation set[46]. Similar images are retrieved from the COCO Caption dataset[47].

hallucinations? To address these inquiries, we aim to decouple the contribution of the visual information in $\boldsymbol{v}$ to the predicted confidence scores. Inspired by Lin *et al.*, we first input the test image $\boldsymbol{v}^\tau$ into the MLLM to predict $\boldsymbol{h}_t(\boldsymbol{v}^\tau, \boldsymbol{x}, \boldsymbol{y}_{<t})$ and greedily decode tokens $[y_1, \ldots, y_n]$[45]. $y_t$'s confidence scores are denoted as the *base scores*. Then at each decoding step $t, t \in \{1, 2, ..., n\}$, the test image is replaced with alternatives $\boldsymbol{v}'$ (either a noised image without valid visual content, or other similar images), and the predicted tokens thus far $[y_1, \ldots, y_{t-1}]$ are concatenated to $\boldsymbol{x}$ to predict a new hidden state $\hat{\boldsymbol{h}}_t(\boldsymbol{v}', \boldsymbol{x}, \boldsymbol{y}_{<t})$ and decode $\hat{y}_t$, as illustrated in Figure 2. The input shift vectors $\Delta\boldsymbol{ve} = VE(\boldsymbol{v}^\tau) - VE(\boldsymbol{v}')$, represent the variation in input visual information, which is anticipated to induce a corresponding output feature shift $\Delta\boldsymbol{h} = \boldsymbol{h}_t(\boldsymbol{v}^\tau, \boldsymbol{x}, \boldsymbol{y}_{<t}) - \hat{\boldsymbol{h}}_t(\boldsymbol{v}', \boldsymbol{x}, \boldsymbol{y}_{<t})$. Therefore, the confidence score distribution shift $(\Delta\boldsymbol{h} \cdot E_c(x')), x' \in \mathcal{V}$ (the subtraction of $\hat{y}_t$'s confidence scores from $y_t$'s) represents the semantics in $\boldsymbol{h}_t$ contributed by the visual information in $\Delta\boldsymbol{ve}$. We examine the impact of the visual information to all predicted tokens' confidence score distributions in the output sentences.

## 3.2. Main findings

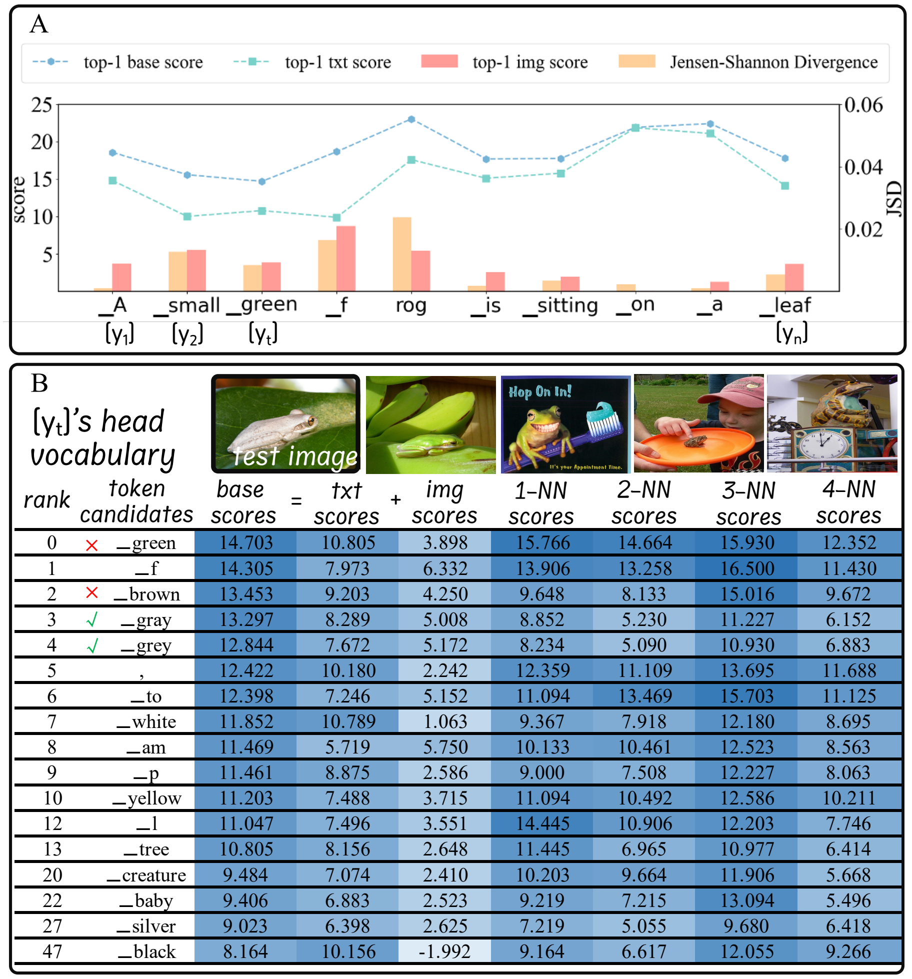
### *3.2.1 MLLMs are aware of accurate visual cues amid hallucination*

To integrally decouple the contribution of the visual branch to MLLMs' predictions, $\Delta\boldsymbol{ve}$ should encapsulate the majority of visual information. To this end, we erase the visual information in the test image $\boldsymbol{v}^\tau$ until it is indistinguishable from Gaussian noise, following the image diffusion[48] process (Note that we only use the diffusion process to add noise to images, rather than employing the Denoising Diffusion Probabilistic Model (DDPM) to generate images or other content, as done in prior work[49–51].),

$$\boldsymbol{v}^d = \sqrt{\bar{\alpha}_T}\boldsymbol{v}^\tau + \sqrt{1 - \bar{\alpha}_T}\epsilon, \tag{3}$$

where $\bar{\alpha}_T = \prod_{i=1}^{T} \alpha_i$ is the cumulative product of the noise schedule parameters $\alpha_i$. $T$ is the diffusion step. $\epsilon$ is Gaussian noise sampled from a normal distribution $\mathcal{N}(0, I)$, $I$ is the identity matrix. Then the MLLM "blindly" predicts a new score distribution (denoted as the txt scores) using $\boldsymbol{v}^d$, devoid of valid visual cues. Thus, the subtraction of the txt scores from the base scores can be interpreted as the contribution of the visual modality (denoted as the img scores).

To quantify the dependency of each predicted token on the visual input, we propose a metric that combines the Jensen-Shannon Divergence (JSD) between the base scores and the txt scores, and the img score of the top-ranked candidate (denoted as the top-1 img score). A high JSD value suggests significance difference between the predicted probabilities corresponding to the base scores and the txt scores. Consequently, the



| rank | token candidates | base scores | = txt scores | + img scores | 1–NN scores | 2–NN scores | 3–NN scores | 4–NN scores |
|---|---|---|---|---|---|---|---|---|
| 0 | ✗ _green | 14.703 | 10.805 | 3.898 | 15.766 | 14.664 | 15.930 | 12.352 |
| 1 | _f | 14.305 | 7.973 | 6.332 | 13.906 | 13.258 | 16.500 | 11.430 |
| 2 | ✗ _brown | 13.453 | 9.203 | 4.250 | 9.648 | 8.133 | 15.016 | 9.672 |
| 3 | ✓ _gray | 13.297 | 8.289 | 5.008 | 8.852 | 5.230 | 11.227 | 6.152 |
| 4 | ✓ _grey | 12.844 | 7.672 | 5.172 | 8.234 | 5.090 | 10.930 | 6.883 |
| 5 | , | 12.422 | 10.180 | 2.242 | 12.359 | 11.109 | 13.695 | 11.688 |
| 6 | _to | 12.398 | 7.246 | 5.152 | 11.094 | 13.469 | 15.703 | 11.125 |
| 7 | _white | 11.852 | 10.789 | 1.063 | 9.367 | 7.918 | 12.180 | 8.695 |
| 8 | _am | 11.469 | 5.719 | 5.750 | 10.133 | 10.461 | 12.523 | 8.563 |
| 9 | _p | 11.461 | 8.875 | 2.586 | 9.000 | 7.508 | 12.227 | 8.063 |
| 10 | _yellow | 11.203 | 7.488 | 3.715 | 11.094 | 10.492 | 12.586 | 10.211 |
| 12 | _l | 11.047 | 7.496 | 3.551 | 14.445 | 10.906 | 12.203 | 7.746 |
| 13 | _tree | 10.805 | 8.156 | 2.648 | 11.445 | 6.965 | 10.977 | 6.414 |
| 20 | _creature | 9.484 | 7.074 | 2.410 | 10.203 | 9.664 | 11.906 | 5.668 |
| 22 | _baby | 9.406 | 6.883 | 2.523 | 9.219 | 7.215 | 13.094 | 5.496 |
| 27 | _silver | 9.023 | 6.398 | 2.625 | 7.219 | 5.055 | 9.680 | 6.418 |
| 47 | _black | 8.164 | 10.156 | -1.992 | 9.164 | 6.617 | 12.055 | 9.266 |

**Figure 3.** Visual hallucination analysis results. (A) the JSD is employed to measure the dependency of each generated token on the visual input. The JSD corresponding to articles and prepositions (such as _a and _on) are close to zero, while the JSD value for the erroneous token _green is significantly higher. (B) LLaVA-1.5 can identify accurate visual cues even amid hallucinations, as the visual information contributes +5.008 confidence scores to the accurate candidate _gray. However, the visual branch also mistakenly supports inaccurate candidates (e.g., +3.898 for _green and +4.250 for _brown). Additionally, images with similar semantics and appearances can induce analogous visual hallucinations. For instance, candidate _green receives high confidence scores (+15.930 and +12.352) in images that do not contain green frogs. JSD: Jensen-Shannon Divergence.

img scores will not be uniformly distributed. This indicates that some, but not all, candidates are substantially influenced (either advocated or suppressed) by the visual branch. Conversely, if both the JSD value and the top-1 img score are close to zero, we assert that the visual information has minimal impact on predictions at the current decoding step, as ablating the input visual information does not result in a significant difference in the predicted probabilities.

The results of JSD and top-1 img scores are presented in Figure 3A. At the third decoding step, where the erroneous token _green is selected, both the JSD and the top-1 img scores are significantly higher than those of the grammatical tokens (e.g., _on and _a) and are not close to zero. Figure 3B presents the predicted base scores, txt scores, and img scores. First, we observe that the predicted img scores are multimodally distributed among the top-ranked candidates. Notably, both accurate and erroneous candidates obtain relatively high img scores, e.g., +5.008 scores for accurate candidate _gray, +3.898 and +4.250 scores for erroneous candidates _green and _brown, respectively. Therefore, the MLLM (LLaVA-1.5) is not entirely disregarding accurate visual cues when it generates hallucinations in this case. On the other hand, this result also reveals that the visual branch

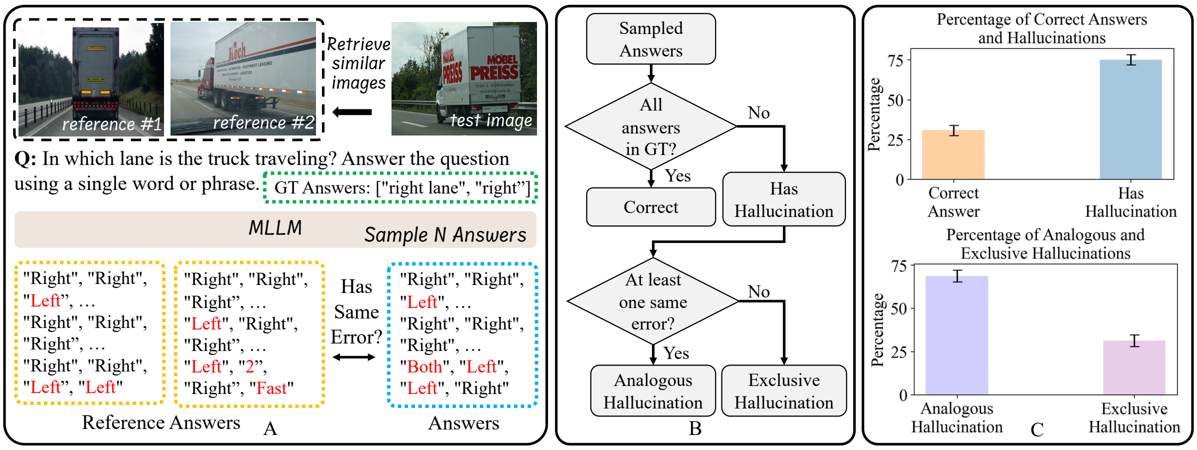

reference #1
reference #2
Retrieve similar images
test image
Q: In which lane is the truck traveling? Answer the question using a single word or phrase.
GT Answers: ["right lane", "right"]
MLLM
Sample N Answers
"Right", "Right", "Left", … "Right", "Right", "Right", … "Right", "Right", "Left", "Left"
"Right", "Right", "Right", … "Left", "Right", "Right", … "Left", "2", "Right", "Fast"
Has Same Error?
"Right", "Right", "Left", … "Right", "Right", "Right", … "Both", "Left", "Left", "Right"
Reference Answers
A
Answers
Sampled Answers
All answers in GT?
No
Yes
Correct
Has Hallucination
At least one same error?
No
Yes
Analogous Hallucination
Exclusive Hallucination
B
Percentage of Correct Answers and Hallucinations
Percentage
75
50
25
0
Correct Answer
Has Hallucination
Percentage of Analogous and Exclusive Hallucinations
Percentage
75
50
25
0
Analogous Hallucination
C
Exclusive Hallucination

**Figure 4.** Quantitative results supporting our hypothesis that similar images might induce analogous hallucinations. (A) Experiment pipeline for investigating the characteristics of visual hallucinations. image-question pairs are randomly selected from the VQAv2 validation set and similar images are retrieved for each test image. We then independently sample $N$ answers for the test image and the retrieved images. (B) We evaluate the correctness of each sampled answer, and analyze the overlap between the incorrect answers predicted using the test image and those derived from the reference images. if all $N$ answers are correct, then the test sample is categorized into "Correct". If at least one incorrect answer exists in the reference answers, the sample is categorized into "Analogous Hallucination". Otherwise, it is categorized into "Exclusive Hallucination". (C) LLaVA-Next exhibits hallucinations in over 70% of the selected samples from VQAv2. Among the samples where hallucinations occur, more than two-thirds demonstrate that similar images induce analogous hallucinations. VQAv2: Visual Question Answering version 2.

of MLLMs may substantially promote hallucinatory token candidates. Second, certain candidates may receive minimal support or even opposition from the visual branch, e.g., candidate _black gets -1.992 img scores. According to previous studies[15,21], the word 'black' in this example reflects erroneous language inherited from superficial syntactical patterns in the training data. In summary, these findings supplement previous studies, indicating that the hallucinatory token candidates are not solely advocated by textual branch of MLLMs.

Regarding the example presented in Figure 3B, employing existing methods[15,21], which adds scaled img scores to base scores, can suppress erroneous language priors. For instance, the base score of the erroneous candidate _black will be reduced. However, this approach will further increase the base scores of the hallucinatory candidates _green and _brown, thereby exacerbating hallucinations. In this work, we aim to eliminate this adverse side effect to achieve better mitigation of visual hallucinations.

### *3.2.2 Analogous visual hallucinations occur among similar images*

Having identified that MLLMs are aware of accurate visual cues even amid visual hallucinations, we then investigate whether similar images tend to induce similar hallucinations with identical textual prompts. To this end, we randomly select 100 image-question pairs from the VQAv2[22] validation set to investigate the patterns of incorrect answers. For each test image, reference images with similar semantics and appearance are retrieved from the COCO Caption[47] dataset utilizing the image retrieval method specified in Section 4.1. We manually verify that each question and its corresponding ground truth answers strictly adhere to both the test image and the retrieved reference images. Ultimately, we obtained 100 test samples and a total of 325 images, with each test image having an average of 2.25 reference images.

LLaVA-Next's answers for each image-question pair are obtained utilizing multi-nominal sampling strategy with fixed decoding parameters. Specifically, $N$ answers are independently sampled for each image (both the test image and references), as illustrated in Figure 4A. Recall that token candidates with higher confidence scores are more likely to be sampled according to Equation (1). Thus, the proportion of each answer's occurrence in the sampled results reflects the model's confidence level in those answers. In practice, we set $N = 20$

and obtain 6,500 generated answers. The Accuracy metric defined using Exact Match (EM) evaluates the correctness of each answer. Next, we assess whether a sample exhibits hallucination and if similar images induce analogous hallucinations following the process outlined in Figure 4B. Specifically, if all $N$ answers fall within the set of ground truth answers, the current answers are considered to be correct. Otherwise, the test sample is classified as Has Hallucination. Furthermore, for samples exhibiting hallucinations, if at least one incorrect answer exists in the set of reference answers (predicted using the reference images), we determine that analogous hallucination has occurred among similar images. For example, the test sample in Figure 4A is classified as Analogous Hallucination, since the hallucinatory answer "Left" exists in the reference answers (both the test image and reference images show a truck on the right lane). Otherwise, we determine that the test sample has exclusive hallucinations.

Experimental Phenomena: The percentage distributions of the defined categories (Correct versus Has Hallucination and Analogous hallucination versus Exclusive Hallucination) are presented in Figure 4C. Notably, LLaVA-Next exhibits hallucinations on 75.2%(±3.2%) of the selected samples. Among the test samples with hallucinatory answers, 68.7%(±3.4%) show that similar images induce analogous visual hallucinations. The results are averaged on five separate experiments with different random seeds.

These experimental results support the following assumption:

- Condition: The input textual prefix to the MLLM is identical.
- Conclusion: Similar images are likely to induce analogous visual hallucinations.

### *3.2.3 Analogous visual hallucinations help discern accurate content*

Having established that similar images can induce analogous hallucinations, we further investigate whether these analogous hallucinations can help identify accurate token candidates by analyzing the differences in the confidence score distributions generated using the test image and the reference images. Figure 3B presents four retrieved images, along with their corresponding confidence scores (in the k-NN scores columns).

Experimental Phenomena: Upon comparing the base scores with the four k-NN scores, a notable observation emerges: The score of the accurate candidate _gray decreases significantly from 13.297 to 7.865 (on average across four references). In contrast, the score changes for visually deceptive candidates (We refer to the inaccurate candidates that are erroneously promoted by the visual branch, as visually deceptive candidates).

(e.g., _green, _brown, and _yellow) and textually deceptive candidates ( We refer to the inaccurate candidates, which are opposed by the visual branch, as textually deceptive candidates). (e.g., _black) are relatively modest. For instance, candidate _green receives 15.390 and 12.352 scores in images without green frogs. Considering token candidates in the head vocabulary illustrated in Figure 3B, the average score variation for accurate candidates, i.e., _gray, _grey, _white, and _silver, is 3.68, while the average score variation for erroneous or ambiguous candidates is 0.60.

Conclusion: The example in Figure 3B demonstrates that, after replacing the input image with another globally similar image under the same text prefix, the confidence score of the accurate token candidates decreases more significantly than that of the hallucinatory candidates.

Inference: Under the same text prefix, token candidates whose confidence scores vary significantly across similar images are more likely to be the correct ones compared to candidates with moderate score changes. Note that we do not deny that candidates with moderate score changes may also represent common visual content in similar images. However, we emphasize that they are more likely to be hallucinatory content than candidates with significant score variations.

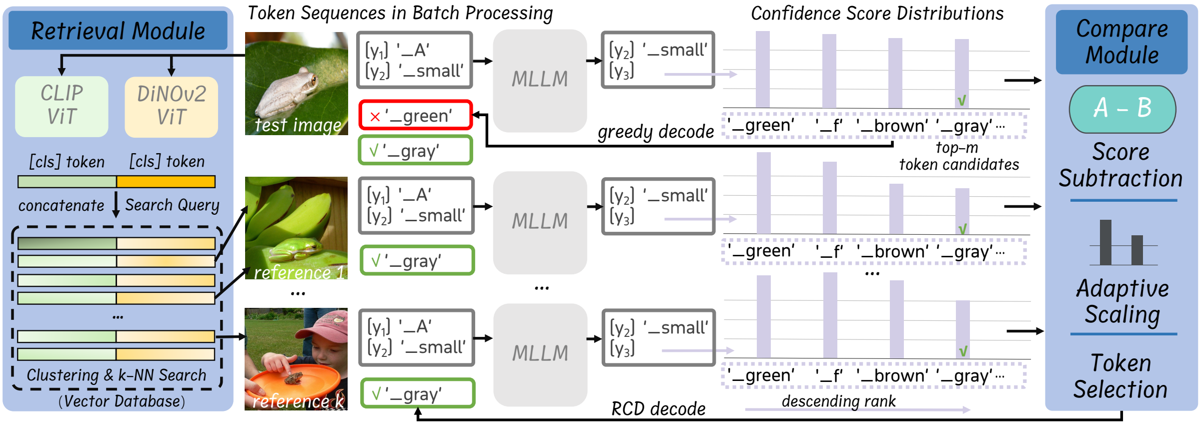


**Figure 5.** Our proposed approach RCD identifies erroneous token candidates and promotes accurate candidates during the inference stage of MLLMs. RCD first retrieves relevant images from a reference database utilizing the retrieval module, and then the MLLM generates distinct predictions for each reference with identical textual prefixes. The predicted confidence scores are then contrasted by RCD's compare module to highlight the accurate candidates and decode the next token. These modules can be seamlessly integrated into existing MLLMs without requiring model retraining. RCD: Retrieval contrastive decoding; MLLMs: multi-modal large language models.

Therefore, the difference in confidence scores predicted by similar images can potentially assist MLLMs in distinguishing between accurate and erroneous token candidates.

## 4. METHOD

Inspired by our observation that MLLMs can recognize accurate visual cues even in the presence of hallucinations, and that analogous hallucinations among similar images can aid MLLMs in identifying accurate visual content, we propose a training-free and plug-and-play method named RCD. This method straightforwardly calibrates the predicted confidence scores of MLLMs, focusing on prioritizing accurate content over hallucinations. Figure 5 presents an overview of RCD, which consists of two key components: the retrieval module, which searches for similar images, and the compare module, which contrasts visual cues. These modules can be seamlessly integrated into existing MLLMs without requiring model retraining.

### 4.1. Retrieve visual references

The reference database is expected to encompass diverse visual concepts, spanning a broad spectrum of common visual content. This is essential for the generalization capability of RCD. Before MLLMs' inference stage, $k$ most relevant visual references are retrieved for each test sample $q$ from the reference database $\mathcal{D}$, based on a similarity measure $\mathcal{F}(\cdot,\cdot)$. Formally,

$$\{r_1,\ldots,r_k|q\} = \arg\max_k\{\mathcal{F}(E_R(s_j),E_R(q))|s_j \in \mathcal{D}\} \tag{4}$$

where $E_R(\cdot)$ is the retriever that embeds the raw inputs into vector representations for image retrieval. $\mathcal{D}[r], r \in \{r_1,\ldots,r_k|q\}$ are the desired references for $q$. Overall, the reference database is analogous to the memory mechanism in the human brain, enabling MLLMs to refer to relevant information during the process of visual recognition. In the following subsections, we provide detailed explanations for the database $\mathcal{D}$, the retriever $E_R(\cdot)$, and the similarity metric $\mathcal{F}(\cdot,\cdot)$.

#### *4.1.1 Reference database*

The reference database is constructed using samples from the COCO Caption dataset. Specifically, both the Karpathy train and rest-val splits[52] are included, totaling around 113,000 samples. This ensures that the database encompasses a wide range of visual concepts. In our database, each sample is stored as a key-value pair for retrieval purposes. The retrieval key is the vector embedding generated by the retriever $E_R(\cdot)$, and the retrieval value is the filename of the image on the disk. The retrieval keys are stored in a vector database format and are clustered to facilitate efficient retrieval. Importantly, all samples that appear in test datasets are excluded from the reference database to prevent information leakage.

#### *4.1.2 Retriever*

The retriever $E_R(\cdot)$ extracts representations from raw inputs (images or text). The representations are then used to measure the similarity between samples. To improve the efficiency and effectiveness of visual reference retrieval, we aim to extract a compact and distinctive representation for each sample that encapsulates its global features. To achieve this, we employ feature extraction models for image and text modalities that have been pre-trained on large-scale datasets. Specifically, We ensemble the CLIP[53] vision transformer (ViT[54]), which excels at encoding semantics, and the self-supervised pre-trained DINOv2[55] ViT, which is proficient in capturing visual details. For image captioning and open-ended VQA, the retriever extracts semantic and appearance features from the visual input $\boldsymbol{v}$ into $E_R(\boldsymbol{v})$ for all test images and reference images. In practice, the predicted $[cls]$ tokens from the penultimate transformer block of the CLIP (Available at https://huggingface.co/openai/clip-vit-large-patch14) and DINOv2 (Available at https://github.com/facebookresearch/dinov2) ViT-L14/336 models are concatenated in the feature dimension to serve as the representation for each image, as illustrated in Figure 5. For yes-or-no binary VQA, we focus on finding visual references that are semantically aligned with the question. The CLIP Transformer is used to extract semantics from the question $\boldsymbol{x}$ into a vector embedding $E_R(\boldsymbol{x})$ for all test samples. For hallucination evaluation benchmarks MME and POPE, we modify the question template before extracting text features, rewriting questions into narratives.

#### *4.1.3 Similarity metric*

The similarity metric $\mathcal{F}(\cdot,\cdot)$ in Equation (4) is implemented using cosine similarity,

$$\mathcal{F}(E_R(s_j), E_R(q)) = \frac{E_R(s_j)\cdot E_R(q)}{\|E_R(s_j)\|\|E_R(q)\|} = \frac{\sum_{i=1}^{n} E_R(s_j)_i E_R(q)_i}{\sqrt{\sum_{i=1}^{n}(E_R(s_j)_i)^2}\sqrt{\sum_{i=1}^{n}(E_R(q)_i)^2}} \tag{5}$$

where $n$ is the size of the feature dimension. Cosine similarity selected for its invariance to the magnitude of vectors, which reduces the impact of numerical discrepancies between image features from different models, ensuring a balanced contribution from each vision encoder. The similarity between the retrieval query $E_R(q)$ (image or text features of all test samples) and all retrieval keys $E_R(s_j)$ (image features of all reference samples $s_j \in \mathcal{D}$) is evaluated using Equation (5). In practice, $E_R(s_j)$ and $E_R(q)$ are first L2 normalized, and maximum inner product search (MIPS) is conducted using FAISS[56], which is a library for vector clustering and similarity search. For image captioning and open-ended VQA, both semantic and appearance-level similarities between the test image and visual references are equally considered, as the normalized $[cls]$ tokens from CLIP and DINOv2 ViTs are concatenated in the feature dimension. For binary VQA, we utilize the cross-modal feature matching capabilities of the CLIP model to retrieve images relevant to the question. On the MME benchmark, the retrieval results are re-ordered according to the BLEU@1[57] score between the question $\boldsymbol{x}$ and the annotated captions of the retrieved images for better performance.

#### *4.1.4 Efficiency*

The FAISS library uses vector clustering and approximate k-nearest neighbor (k-NN) algorithms for acceleration. We evaluate the efficiency of the retrieval process on a single NVIDIA 3090 GPU. When retrieving 32 visual references for each query, calculating $E_R(\boldsymbol{v})$ takes up around 40 ms, and the retrieval process takes under 5 ms (averaged on 500 queries). It is important to note that RCD does not increase the input token sequence length of MLLMs, and the additional token sequences can be stacked for batch processing to further reduce computational latency. A detailed efficiency analysis is provided in Section 5.5.4.

### 4.2. Compare visual concepts

In Section 3.2.2 and Section 3.2.3, we demonstrate that similar images can induce analogous hallucinations, and infer that, after replacing the test image with globally similar ones (under identical text prefixes), token candidates exhibiting more significant confidence score decline (i.e., BaseScores-kNNScores in Figure 3B) are more likely to be the correct ones. Therefore, contrasting the confidence scores predicted by similar images can help MLLMs distinguish accurate visual cues and mitigate visual hallucinations. Specifically, the MLLM first generates $k + 2$ distinct confidence scores (referred to as logits in Equations (6) and (7) for each token candidate $x_i \in \mathcal{V}$. These scores are derived from one test image $\boldsymbol{v}^\tau$, one diffused image $\boldsymbol{v}^d$, and $k$ retrieved images $\{\boldsymbol{v}^{NN}\}_k$, all with identical textual prefix $\boldsymbol{x} + \boldsymbol{y}_{<t}$, as illustrated in Figure 5. Then, $x_i$'s confidence score predicted by the test image $\boldsymbol{v}^\tau$ is contrasted to the scores predicted by the $k + 1$ references,

$$\begin{aligned} logits(x_i|\boldsymbol{x}, \boldsymbol{y}_{<t}, \boldsymbol{v}^\tau, \boldsymbol{v}^d, \{\boldsymbol{v}^{NN}\}_k) &= (\alpha_\tau + \alpha_d^t + \alpha_{NN}^t)\, logits(x_i|\boldsymbol{x}, \boldsymbol{y}_{<t}, \boldsymbol{v}^\tau) \\ &- \frac{\alpha_{NN}^t}{k} \sum_{j=1}^{k} logits(x_i|\boldsymbol{x}, \boldsymbol{y}_{<t}, \boldsymbol{v}_j^{NN}) - \alpha_d^t logits(x_i|\boldsymbol{x}, \boldsymbol{y}_{<t}, \boldsymbol{v}^d) \end{aligned} \quad (6)$$

where the subtraction operator highlights the difference between the test image and the visual references. This approach prioritizes candidates with significant score variations across similar images, as they are more likely to be the correct ones, over candidates with moderate score changes, which may be the analogous visual hallucinations across similar images. The subtracted scores $logits(x_i|\boldsymbol{x}, \boldsymbol{y}_{<t}, \boldsymbol{v}^\tau) - \frac{1}{k}\sum_{j=1}^{k} logits(x_i|\boldsymbol{x}, \boldsymbol{y}_{<t}, \boldsymbol{v}_j^{NN})$ and $logits(x_i|\boldsymbol{x}, \boldsymbol{y}_{<t}, \boldsymbol{v}^\tau) - \alpha_d^t logits(x_i|\boldsymbol{x}, \boldsymbol{y}_{<t}, \boldsymbol{v}^d)$ are then scaled and added to the confidence scores corresponding to the test image $logits(x_i|\boldsymbol{x}, \boldsymbol{y}_{<t}, \boldsymbol{v}^\tau)$. This ensures that the output of MLLMs remains unchanged when the difference in predicted confidence score distributions is close to zero (in circumstances when the MLLM determines that the retrieved images are almost identical to the test image). The outcome of Equation (6) yields probabilities after the softmax operator [Equation (1)], and these predicted probabilities are subsequently used for token selection. The selected token is appended to all sequences to ensure that the textual prefix remains identical at each decoding step. This confidence score calibration process is applied to MLLMs throughout the output sentences.

#### *4.2.1 Adaptive scaling strategy*

The influence of $\{\boldsymbol{v}^{NN}\}_k$ and $\boldsymbol{v}^d$ should be regulated in order to coordinate the effect in mitigating hallucinations originating from both visual or textual branches of MLLMs. Specifically, in cases where MLLMs exhibit uncertainty[17] regarding the recognized visual cues, i.e., the confidence scores, $\{l_{i,t}^\delta\}_{i=0}^{m-1}$, exhibit multimodal distribution, which are calculated by

$$l_{i,t}^\delta = logits(x_i|\boldsymbol{x}, \boldsymbol{y}_{<t}, \boldsymbol{v}^\tau) - logits(x_i|\boldsymbol{x}, \boldsymbol{y}_{<t}, \boldsymbol{v}^d) \quad \text{s.t.} \quad x_i \in \mathcal{V}_{head}^m \quad (7)$$

then the coefficient $\alpha_d^t$ is reduced while $\alpha_{NN}^t$ is increased at the current decoding step $t$. This adjustment ensures that the token candidates, which are erroneously promoted by the visual branch, are not further endorsed. Formally,

$$\alpha_d^t = \beta_d \exp\left(\max(\text{softmax}(\{l_{i,t}^\delta\}_{i=0}^{m-1}))\right) \quad (8)$$

$$\alpha_{NN}^t = \beta_{NN} \exp\left(1 - \max(\text{softmax}(\{l_{i,t}^\delta\}_{i=0}^{m-1}))\right) \quad (9)$$

Following the adaptive plausibility constraint[41], RCD only considers $x_i$ that are in the head vocabulary $\mathcal{V}_{head}^m$, which consists of $m$ top-ranked candidates that are selected based on the confidence scores predicted by $\boldsymbol{v}^\tau$ (i.e., the base scores). In practice, we set a cut-off value as the base score of the $m^{th}$-ranked candidate. Confidence scores lower than this cut-off value are set to $-inf$. $m$ is set at 50 by default. $\alpha_\tau$, $\beta_d$ and $\beta_{NN}$ are hyperparameters, which are by default set at 1.0, 0.1, and 0.1, respectively.

### 4.3. Summary

Algorithm 1 provides an overview of our proposed method RCD. First, the search query $q$ is flexibly configured according to the different types of visual language tasks. Next, the $search_{kNN}$ function [Equation (4)] is used

**Algorithm 1** Our retrospect-then-compare paradigm

**Input**: a test image $\boldsymbol{v}^\tau$, input textual prefix $\boldsymbol{x}$, output token list $\boldsymbol{y} = []$, the reference database $\mathcal{D}$, the retriever $E_R(\cdot)$

**Arguments**: number of references $k$, diffusion step $ds$, head vocabulary size $m$, hyper-parameters $\alpha_\tau$, $\beta_d$ and $\beta_{NN}$

**Output**: output token list $\boldsymbol{y}$

```
if Binary VQA task then
    Let $q = \boldsymbol{x}$.
else
    Let $q = \boldsymbol{v}^\tau$. # image captioning or open-ended VQA
end if
$\{\boldsymbol{v}^{NN}\}_k = search_{kNN}(\mathcal{D}, E_R(\cdot), q, k)$.
$\boldsymbol{v}^d = diffuse(\boldsymbol{v}^\tau, ds)$
$t = 0$ # decoding step
$stop\_condition = False$
while not $stop\_condition$ do
    $l_{v^\tau} = logits(x|\boldsymbol{x}, \boldsymbol{y}_{<t}, \boldsymbol{v}^\tau) = LMHead(LLM_\theta([VE(\boldsymbol{v}^\tau);TE([\boldsymbol{x};\boldsymbol{y}_{<t}])]))$
    $l_{v^d} = logits(x|\boldsymbol{x}, \boldsymbol{y}_{<t}, \boldsymbol{v}^d) = LMHead(LLM_\theta([VE(\boldsymbol{v}^d);TE([\boldsymbol{x};\boldsymbol{y}_{<t}])]))$
    $l_{v_j^{NN}} = logits(x|\boldsymbol{x}, \boldsymbol{y}_{<t}, \boldsymbol{v}_j^{NN}) = LMHead(LLM_\theta([VE(\boldsymbol{v}_j^{NN});TE([\boldsymbol{x};\boldsymbol{y}_{<t}])]))$, where j = 1,2,..., k
    $\mathcal{V}_{head}^m = \arg\max_m(l_{v^\tau})$
    $\alpha_d^t, \alpha_{NN}^t = AS(l_{v^\tau}, l_{v^d}, \beta_d, \beta_{NN}, \mathcal{V}_{head}^m)$
    $logits(x|\boldsymbol{x}, \boldsymbol{y}_{<t}, \boldsymbol{v}^\tau, \boldsymbol{v}^d, \{\boldsymbol{v}^{NN}\}_k) = contrast(\alpha_d^t, \alpha_{NN}^t, \alpha_\tau, l_{v^\tau}, \{l_{v_j^{NN}}\}_{j=1}^k, l_{v^d})$
    $t = t + 1$
    $y_t = decode(logits(x|\boldsymbol{x}, \boldsymbol{y}_{<t}, \boldsymbol{v}^\tau, \boldsymbol{v}^d, \{\boldsymbol{v}^{NN}\}_k))$
    $\boldsymbol{y}_{<t+1} = [\boldsymbol{y}_{<t}; y_t]$
    $stop\_condition = check\_stop\_condition(stop\_condition)$
end while
return $\boldsymbol{y}$
```

to find $k$ most relevant images for each test sample. Additionally, we add noise to the test image [Equation (3)] as an extra reference to address erroneous language bias, following the approach in VCD[15]. During the inference stage, $k + 2$ distinct confidence score distributions are independently predicted by the language model. Subsequently, the Adaptive Scaling strategy (AS) regulates the influence of $\{\boldsymbol{v}^{NN}\}_k$ and $\boldsymbol{v}^d$, The visual concept comparison step [the $contrast(\cdot)$ function, Equation (6)] then adjusts the predicted confidence score distribution. Finally, token candidates are selected based on the modified confidence scores through the $decode(\cdot)$ function. This confidence score calibration process is repeated until the stop condition is met. In summary, we develop a reference database that functions similarly to the memory mechanism in the human brain, allowing MLLMs to access relevant information during the visual recognition process. Our proposed confidence score calibration mechanism is designed to distinguish accurate visual content from hallucinations. Notably, RCD addresses visual hallucinations originating from both the visual and textual branches of MLLMs, overcoming the limitations of previous methods that could only suppress erroneous language priors from the text decoder. Additionally, RCD can be easily integrated into MLLMs without requiring model retraining.

## 5. RESULT

### 5.1. Experimental settings

*5.1.1 Datasets and metrics*

We use publicly available hallucination benchmarks POPE [25], (the MLLM Evaluation benchmark) MME[24], WHOOPS[26], and LLaVA-Bench-in-the-wild[27] to validate the effectiveness of our proposed method. POPE is designed to assess hallucinations at the object level, while MME evaluates hallucinations at both the object level and the visual attribute level. In these benchmarks, MLLMs are prompted to answer the questions using a single word or phrase. The evaluation metrics for POPE include Accuracy and F1 score, based on whether the prediction contains the annotated answer (either "yes" or "no"). For MME, the evaluation metric is the combined score of Accuracy and Accuracy+. Given that MLLMs encode rich real-world conventions and may embed certain superficial syntactical patterns in model parameters[17], we hypothesize that images containing counterintuitive visual cues (visual content that contradicts common sense) can highlight the problem of visual hallucinations. Therefore, we conduct quantitative experiments using the WHOOPS benchmark, where MLLMs are instructed to describe the counterintuitive images in a single sentence. For the image captioning task, automatic evaluation metrics such as BLEU[57], METEOR[58], CIDEr[59], and SPICE[60] are employed. Additionally, we conduct qualitative experiments on the "detailed description" subset of the LLaVA-Bench-in-the-wild benchmark (LLaVA-W), which encompasses a diverse array of image styles, including real photographs, text-rich images, and sketches.

We further evaluate the effectiveness of RCD in enhancing the MLLMs' capability in real-world scene comprehension and reasoning. To this end, we develop a challenging image captioning and VQA test set, which contains images of complex real-world traffic situations and accidents. Images are sourced from copyright-free websites and high-quality images are retained through manual inspection. For the VQA task, human volunteers are enlisted to write one question for each image based on our devised question categories. Subsequently, we invite 10 volunteers with over 3 years of driving experience to annotate 10 answers for each question. Finally, we manually check all the answers to ensure consistency. In total, 300 image captioning test samples with 1,500 annotated captions, and 200 VQA test samples with 2,000 annotated answers are obtained. For the image captioning task, we use a multi-modal supermodel ( gpt-4o-2024-08-06 model, https://platform.openai.com/docs/models/gpt-4o) to generate captions for each image, and then manually verify and revise these captions to ensure their accuracy.

For the VQA task, we use the EM metric for quantitative evaluation, following the implementation as in the VQAv2[22] benchmark. We carefully devise six categories of questions that are essential for comprehending traffic scenes, encompassing basic visual recognition tasks, logical reasoning tasks, and knowledge-intensive tasks, as detailed below:

- Object Recognition: We design questions regarding the types of objects in the scene (such as the categories and sizes of obstacles on the road), their quantities, and the distances to objects. The proportion of this type of question accounts for 24.2%.
- Traffic Sign Recognition: We design optical character recognition (OCR) questions based on the traffic signs, such as asking about the speed limit, the meaning of prohibition signs, and the names and distances of destinations. We also include questions about traffic lights. The proportion of this type of question accounts for 26.8%.
- Lane Identification: This task involves questions about the number of lanes on the road, the directions of the lanes, and which lane should be selected under specific conditions. The proportion of this type of question accounts for 5.6%.
- Traffic Police Hand Signal Recognition: This is a knowledge-intensive task. The questions ask about the meaning of the traffic police gestures in each image. The proportion of this type of question accounts for 19.7%.
- Reasoning and Forecasting: This task requires the MLLM to observe the information in the image, combined with common sense, and deduce the reasons behind a certain state or action depicted in the picture,

**Table 1. Results on all perception sub-tasks of the MME benchmark**

| Model | Decoding | Color | Count | Existence | Position | Posters | Celebrity | Scene | Landmark | Artwork | OCR | Total |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| LLaVA-Next | greedy | 185.0 | 158.3 | 185.0 | 116.7 | 145.9 | 137.1 | 159.3 | 134.3 | 120.2 | 162.5 | 1504.3 |
| | +*DoLa* | 185.0 | **158.3** | 185.0 | **116.7** | 143.9 | 135.0 | 158.5 | 132.8 | 117.5 | 162.5 | 1490.2 |
| | +*VCD* | **190.0** | 155.0 | 185.0 | 115.0 | 145.9 | 134.1 | 155.5 | 138.3 | 124.3 | 162.5 | 1505.6 |
| | +*Ours* | 185.0 | 155.0 | **185.0** | 113.3 | **155.8** | **152.6** | **161.5** | **145.5** | **124.5** | **162.5** | **1540.7** |
| LLaVA-1.5 | greedy | 155.0 | **158.3** | 195.0 | 123.3 | 129.6 | 132.6 | 155.0 | 163.5 | 121.0 | 125.0 | 1458.9 |
| | +*DoLa* | 153.3 | 158.3 | 195.0 | 123.3 | 127.6 | 130.9 | 154.8 | 162.8 | 122.3 | 122.5 | 1450.7 |
| | +*VCD* | 148.3 | 158.3 | 190.0 | 126.7 | 136.7 | 147.4 | 148.8 | **166.0** | 122.5 | 130.0 | 1474.7 |
| | +*Ours* | **165.0** | 153.3 | **195.0** | **128.3** | **141.8** | **150.3** | **157.3** | 161.8 | **122.8** | **140.0** | **1515.6** |
| InstructBLIP | greedy | 120.0 | 60.0 | 185.0 | 50.0 | 142.9 | 81.8 | 160.0 | 160.0 | 92.0 | 65.0 | 1116.6 |
| | +*DoLa* | 120.0 | 60.0 | 185.0 | 50.0 | 142.9 | 80.9 | 160.0 | 160.0 | 92.2 | 65.0 | 1116.0 |
| | +*VCD* | 123.3 | 60.0 | **185.0** | 53.3 | **151.7** | **94.1** | 156.5 | **161.3** | **99.3** | 95.0 | 1179.5 |
| | +*Ours* | **153.3** | **78.3** | 180.0 | **58.3** | 140.5 | 71.2 | **163.8** | 158.3 | 94.3 | **95.0** | **1193.0** |

We report the officially defined metric that combines Accuracy and Accuracy+ (larger is better). Our proposed method improves the perception competencies for three MLLMs. Bold font indicates that the method in the corresponding row achieves the highest performance for the corresponding MLLM on the given task. MME: OCR: optical character recognition; MLLMs: multi-modal large language models.

or to predict what actions a car or a pedestrian might take in the near future. The proportion of this type of question accounts for 11.1%.

- Driving Maneuver: This task involves questions about which driving behaviors are permitted or prohibited in the scenarios depicted in the image, as well as what actions should be taken to avoid the risk factors. These questions correspond to images that feature different road conditions (dry or icy), different traffic situations (congested or clear), and various traffic signal indications. The proportion of this type of question accounts for 12.6%.

Among the test samples, images depicting traffic accidents (e.g., collisions and fire) and significant risk factors (e.g., slippery roads, obstacles, and imminent collisions) account for 59.6%, nighttime images account for 17.9%, and images from the driver's first-person perspective account for 25.8%. This test set is expected to comprehensively evaluate the capability of MLLMs in comprehending real-world scenes.

#### *5.1.2 MLLM baselines*

Three state-of-the-art MLLMs are selected to implement our method, including InstructBLIP[3], LLaVA-1.5[2], and LLaVA-Next (stronger)[23]. These three MLLMs integrate CLIP Vision Transformer as their visual encoder. As for the cross-modal connector, LLaVA-1.5 and LLaVA-Next use a two-layer MLP, and InstructBLIP uses a Q-former[6] with textual query. We use the 7B version (with around 7-8 billion learnable parameters) of the MLLMs, with Vicuna-7B[42] as the language decoder for LLaVA-1.5 and InstructBLIP, and LLaMA3-8B (Available at https://github.com/meta-llama/llama3/blob/main/MODEL_CARD.md) for LLaVA-Next.

#### *5.1.3 Implementation details*

All experiments are in zero-shot manner, and greedy decoding serves as the baseline strategy for reproducibility. All random seeds are set at 42, ensuring that repeated trials would yield identical results. We run all experiments on two NVIDIA 3090 GPUs. We use PyTorch[61] and the Huggingface Transformers library (https://github.com/huggingface/transformers) to implement our method. We compare RCD with two advanced decoding strategies VCD[15] and DoLa[62], using their official code implementations. VCD contrasts the confidence scores of the test image with the scores of a noised image, targeting at reducing erroneous language priors. For VCD, we follow the official hyper-parameters on the MME and POPE benchmarks. DoLA aims to amplify the knowledge injected by deeper layers by contrasting the confidence scores predicted by the language model's final hidden state with those predicted by its shallow layers. For DoLa, we follow its dynamic premature layer selection strategy.

### 5.2. Hallucination benchmark results

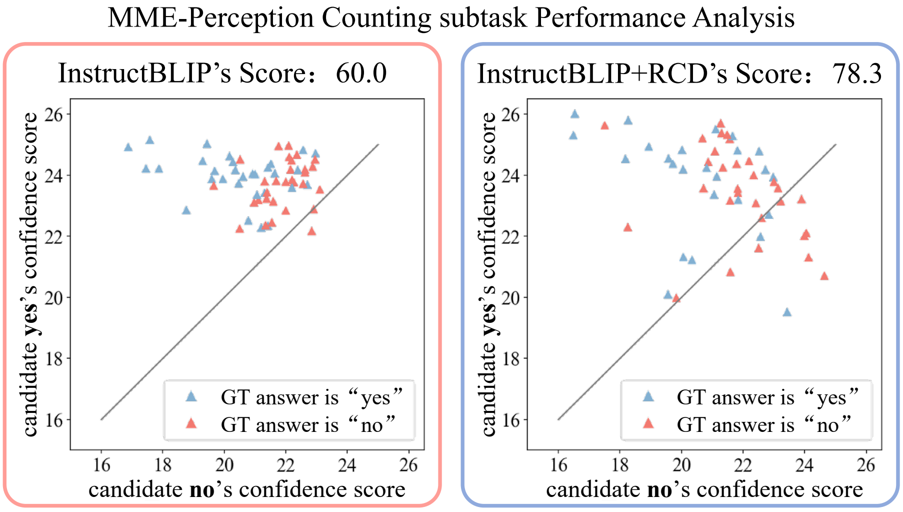


**Figure 6.** Analysis of RCD's impact on the MME benchmark. Our proposed RCD reduces the answer uncertainty of InstructBLIP, and decreases the number of false positive samples. RCD: Retrieval contrastive decoding; MME:

### *5.2.1 Results on MME*

Experimental results on the perception subtasks of MME are presented in Table 1. Without additional training, RCD significantly improves the total score for LLaVA-Next (+36.4 scores), LLaVA-1.5 (+56.6 scores), and InstructBLIP (+76.3 scores), surpassing current advanced decoding strategies DoLa and VCD by 64.1 and 29.8 scores on average, respectively. In particular, RCD effectively mitigates visual hallucinations in color-related perception tasks for LLaVA-1.5 (+10 scores) and InstructBLIP (+33.3 scores). RCD also improves the OCR capability for LLaVA-1.5 (+15 scores) and InstructBLIP (+30 scores).

We further investigate how RCD enhances model performance on the MME benchmark and present the results in Figure 6. We plot the predicted confidence scores corresponding to token candidates _yes and _no ( We observe that in all test results, candidates _yes and _no consistently rank as the top two positions in the vocabulary). on the vertical and horizontal axis, respectively. Each test sample corresponds to a single triangle mark colored in blue (the samples' ground truth answer is _yes) or red (the ground truth answer is _no). If a blue triangle mark is below the $y = x$ line, or a red triangle mark is above the $y = x$ line, the answer is incorrect. The main observations are summarized as follows:

- Without visual references, InstructBLIP has equal confidence, i.e., high uncertainty, to the positive and negative answers. As illustrated in Figure 6, a number of test samples in the Count sub-task are close to the $y = x$ line.
- Without visual references, InstructBLIP tends to answer "yes" for all questions. As shown in Figure 6, almost all samples are located above the $y = x$ line.

Figure 6 illustrates that after integrating RCD, there are fewer red triangle marks above the $y = x$ line. This indicates that the number of false positive answers is reduced. Besides, the sample points demonstrate more dispersed distribution above or under the $y = x$ line with RCD, indicating that the level of uncertainty is reduced.

### *5.2.2 Results on POPE*

The averaged Accuracy and F1 score across the random, popular, and adversarial splits of POPE are presented in Table 2. On three subsets of POPE, RCD boosts the overall Accuracy of LLaVA-Next (+1.98 on average), LLaVA-1.5 (+0.25 on average), and InstructBLIP (+0.91 on average), outperforming VCD and DoLA in varying

**Table 2. Results on the POPE benchmark**

| Model | Decoding | COCO | | AOKVQA | | GQA | |
|---|---|---|---|---|---|---|---|
| | | Acc. ↑ | F1 ↑ | Acc. ↑ | F1 ↑ | Acc. ↑ | F1 ↑ |
| LLaVA-Next | greedy | 83.93 | 81.33 | 84.16 | 83.23 | 82.84 | 81.94 |
| | +*DoLa* | 83.77 | 81.09 | 84.12 | 83.17 | 82.63 | 81.64 |
| | +*VCD* | 86.51 | 85.25 | 85.74 | 85.55 | 84.08 | 83.90 |
| | +*Ours* | **86.73** | **85.65** | **85.93** | **85.85** | **84.21** | **84.11** |
| LLaVA-1.5 | greedy | 85.56 | 84.11 | 84.32 | 84.40 | 84.73 | 84.84 |
| | +*DoLa* | 85.38 | 83.86 | 84.30 | 84.33 | 84.71 | 84.76 |
| | +*VCD* | 85.59 | **85.53** | 82.07 | 83.39 | 82.17 | 82.87 |
| | +*Ours* | **85.80** | 84.58 | **84.56** | **84.91** | **85.01** | **85.38** |
| InstructBLIP | greedy | 85.14 | 84.45 | 81.73 | 83.18 | 80.58 | 82.03 |
| | +*DoLa* | **85.21** | **84.54** | 81.74 | 83.22 | 80.56 | 82.03 |
| | +*VCD* | 84.42 | 83.62 | 81.50 | 82.78 | 80.90 | 82.05 |
| | +*Ours* | 85.12 | 84.37 | **83.30** | **83.88** | **81.76** | **82.15** |

Our proposed method RCD demonstrates superior performance to existing advanced de-coding strategies (VCD and DoLa) on the MSCOCO, AOKVQA and GQA subsets. Bold font indicates that the method in the corresponding row achieves the highest perfor-mance for the corresponding MLLM on the given task. POPE: Polling-based object prob-ing evaluation; RCD: retrieval contrastive decoding; VCD: visual contrastive decoding.

**Table 3. Results on WHOOPS image captioning benchmark**

| Model | Method | B4 ↑ | M ↑ | C ↑ | S ↑ |
|---|---|---|---|---|---|
| LLaVA-Next | greedy | 22.7 | 24.5 | 77.7 | 17.1 |
| | +*DoLa* | 22.6 | 24.4 | 77.9 | 17.0 |
| | +*VCD* | 23.1 | 25.1 | 81.2 | 17.2 |
| | +*RCD* | **23.6** | **25.8** | **83.5** | **17.5** |
| LLaVA-1.5 | greedy | 19.7 | 25.6 | 67.9 | 17.3 |
| | +*DoLa* | 19.9 | 25.6 | 67.8 | 17.4 |
| | +*VCD* | 19.1 | 25.4 | 69.1 | 17.3 |
| | +*RCD* | **20.0** | **26.3** | **75.5** | **17.8** |
| InstructBLIP | greedy | 24.9 | 26.5 | 87.3 | 18.2 |
| | +*DoLa* | 24.8 | 26.5 | 87.4 | 18.2 |
| | +*VCD* | 25.5 | 27.0 | 89.2 | 18.2 |
| | +*RCD* | **25.7** | **27.1** | **90.6** | **18.6** |

RCD boosts the overall performance for LLaVA-1.5 and InstructBLIP. B4, M, C, and S denotes the Bleu@4, METEOR, CIDEr, and SPICE metrics, respectively. Bold font indicates that the method in the corresponding row achieves the highest performance for the corresponding MLLM on the given task. RCD: Retrieval contrastive decoding.

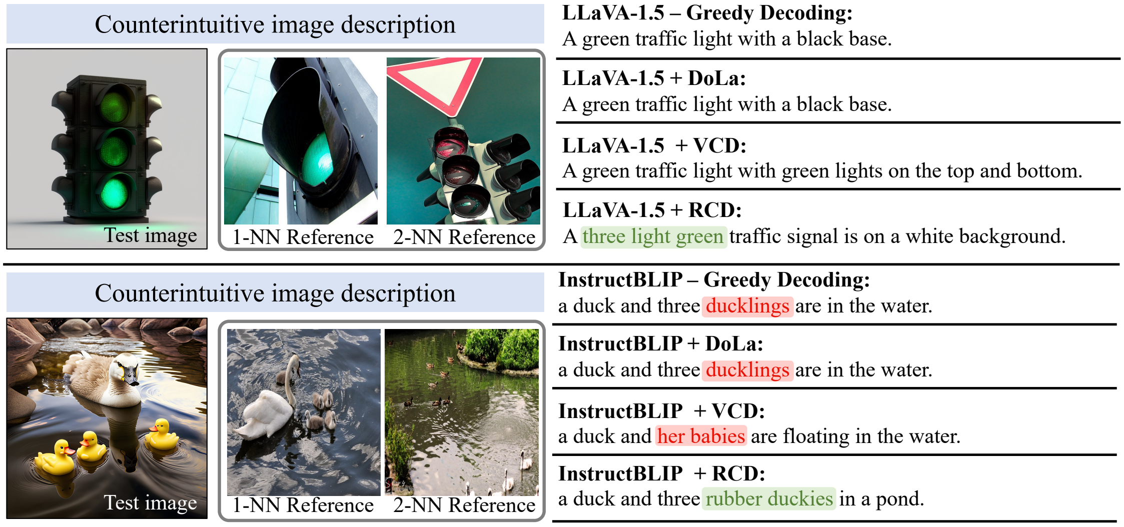
degrees.

### *5.2.3 Results on WHOOPS*

Quantitative results are presented in Table 3. RCD significantly enhances the image description accuracy of LLaVA-1.5 and InstructBLIP. For instance, RCD improves the CIDEr score of LLaVA-1.5 and LLaVA-Next by 7.6 and 5.8, respectively, surpassing existing methods VCD and DoLa. These experimental results validate the effectiveness of RCD in image captioning task and demonstrate its ability to generalize to counterintuitive images. The examples in Figure 7 illustrate the effectiveness of RCD in enhancing MLLMs' comprehension of counterintuitive visual cues. For instance, RCD enables LLaVA-1.5 to accurately identify a traffic signal with three green lights and assists InstructBLIP in correctly distinguishing rubber ducks from real ducks. These results demonstrate that RCD has strong generalization capabilities, enabling precise visual understanding in scenarios involving counterintuitive visual content.

### *5.2.4 Results on LLaVA-W*

The previous sections quantitatively evaluate the effectiveness of RCD. This section qualitatively assesses RCD's performance by presenting examples from the image detailed captioning task on the LLaVA-Bench-in-the-



**LLaVA-1.5 – Greedy Decoding:**
A green traffic light with a black base.

**LLaVA-1.5 + DoLa:**
A green traffic light with a black base.

**LLaVA-1.5 + VCD:**
A green traffic light with green lights on the top and bottom.

**LLaVA-1.5 + RCD:**
A three light green traffic signal is on a white background.

**InstructBLIP – Greedy Decoding:**
a duck and three ducklings are in the water.

**InstructBLIP + DoLa:**
a duck and three ducklings are in the water.

**InstructBLIP + VCD:**
a duck and her babies are floating in the water.

**InstructBLIP + RCD:**
a duck and three rubber duckies in a pond.

**Figure 7.** RCD demonstrates robust generalization capabilities to counterintuitive images, thereby enhancing the accuracy of MLLMs' visual understanding. Correct and hallucinatory content are highlighted in green and red, respectively. RCD: Retrieval contrastive decoding.

wild (LLaVA-W) benchmark, offering a more intuitive analysis of its capabilities. The examples illustrated in Figure 8 demonstrate that RCD effectively mitigates hallucinatory content in detailed descriptions for both in-door and out-door scenes that contain diverse objects, outperforming existing methods DoLa and VCD. For instance, RCD enables LLaVA-1.5 to accurately discern the spatial relationships among multiple instances in a refrigerator, and helps InstructBLIP correctly distinguish streetlights from traffic lights. These examples demonstrate that RCD effectively mitigates various categories of visual hallucination, thereby enhancing MLLMs' capabilities in comprehending complex real-world scenes.

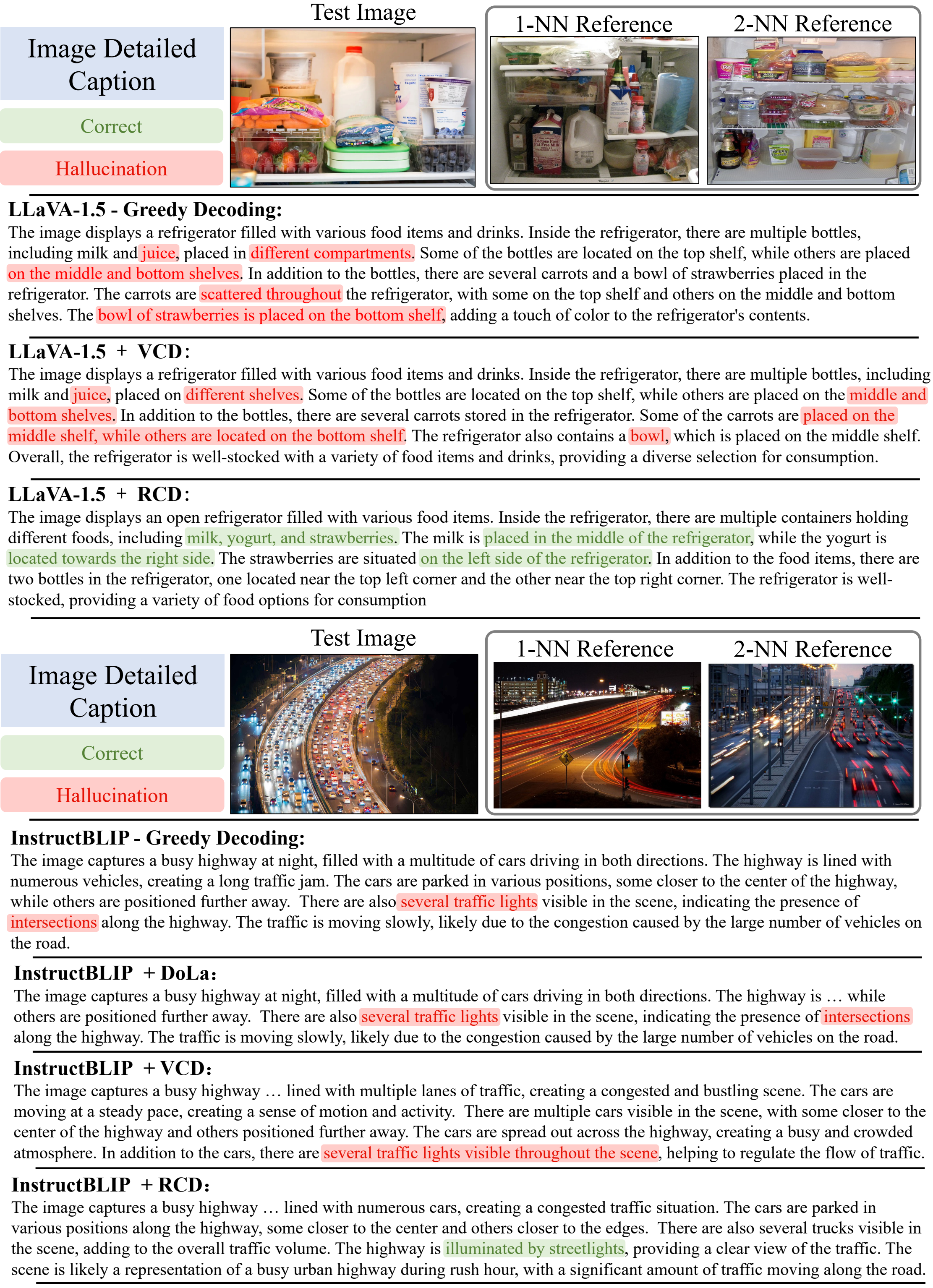
## 5.3. Real world application

### *5.3.1 Quantitative results*

Table 4 demonstrates that RCD enhances the capability of MLLMs to comprehend real-world scenarios. RCD leads to notable improvements in the LLaVA model family for both image captioning and VQA. Specifically, RCD increases the CIDEr score by 3.0 for LLaVA-Next and 1.5 for LLaVA-1.5 in image captioning, and improves the EM metric by 2.4 and 1.2 for LLaVA-Next and LLaVA-1.5 in VQA, respectively. However, the performance gains for InstructBLIP are less pronounced, with only a 0.5 increase in CIDEr scores and a 0.2 increase in EM scores. This may be attributed to the loss of information in the visual token compression process executed by the Q-former[6] model. These results also demonstrate that RCD consistently outperforms DoLA and VCD.

### *5.3.2 Qualitative results*

Qualitative experimental results on our proposed VQA benchmark are presented in Figure 9. Our method demonstrates higher answer accuracy than DoLa and VCD. Specifically, our approach help MLLMs accurately recognize object categories (e.g., bags and cones on the road), and interpret traffic signs (e.g., a yellow triangle sign indicating children crossing) Additionally, RCD enhances the MLLM's reasoning capabilities, such as inferring appropriate driving behaviors (e.g., proceeding at a green light intersection), and predicting potential risks in the scene (e.g., possible collision with a motorcycle on the right side). It is noteworthy that RCD also exhibits robust performance on hazy and low-light images. For particularly challenging questions, such as selecting a lane based on specific requirements, all MLLMs and methods struggle. This suggests that current MLLMs still have considerable room for improvement in accurately comprehending real-world scenarios.



**LLaVA-1.5 - Greedy Decoding:**

The image displays a refrigerator filled with various food items and drinks. Inside the refrigerator, there are multiple bottles, including milk and juice, placed in different compartments. Some of the bottles are located on the top shelf, while others are placed on the middle and bottom shelves. In addition to the bottles, there are several carrots and a bowl of strawberries placed in the refrigerator. The carrots are scattered throughout the refrigerator, with some on the top shelf and others on the middle and bottom shelves. The bowl of strawberries is placed on the bottom shelf, adding a touch of color to the refrigerator's contents.

**LLaVA-1.5 + VCD:**

The image displays a refrigerator filled with various food items and drinks. Inside the refrigerator, there are multiple bottles, including milk and juice, placed on different shelves. Some of the bottles are located on the top shelf, while others are placed on the middle and bottom shelves. In addition to the bottles, there are several carrots stored in the refrigerator. Some of the carrots are placed on the middle shelf, while others are located on the bottom shelf. The refrigerator also contains a bowl, which is placed on the middle shelf. Overall, the refrigerator is well-stocked with a variety of food items and drinks, providing a diverse selection for consumption.

**LLaVA-1.5 + RCD:**

The image displays an open refrigerator filled with various food items. Inside the refrigerator, there are multiple containers holding different foods, including milk, yogurt, and strawberries. The milk is placed in the middle of the refrigerator, while the yogurt is located towards the right side. The strawberries are situated on the left side of the refrigerator. In addition to the food items, there are two bottles in the refrigerator, one located near the top left corner and the other near the top right corner. The refrigerator is well-stocked, providing a variety of food options for consumption

**InstructBLIP - Greedy Decoding:**

The image captures a busy highway at night, filled with a multitude of cars driving in both directions. The highway is lined with numerous vehicles, creating a long traffic jam. The cars are parked in various positions, some closer to the center of the highway, while others are positioned further away. There are also several traffic lights visible in the scene, indicating the presence of intersections along the highway. The traffic is moving slowly, likely due to the congestion caused by the large number of vehicles on the road.

**InstructBLIP + DoLa:**

The image captures a busy highway at night, filled with a multitude of cars driving in both directions. The highway is … while others are positioned further away. There are also several traffic lights visible in the scene, indicating the presence of intersections along the highway. The traffic is moving slowly, likely due to the congestion caused by the large number of vehicles on the road.

**InstructBLIP + VCD:**

The image captures a busy highway … lined with multiple lanes of traffic, creating a congested and bustling scene. The cars are moving at a steady pace, creating a sense of motion and activity. There are multiple cars visible in the scene, with some closer to the center of the highway and others positioned further away. The cars are spread out across the highway, creating a busy and crowded atmosphere. In addition to the cars, there are several traffic lights visible throughout the scene, helping to regulate the flow of traffic.

**InstructBLIP + RCD:**

The image captures a busy highway … lined with numerous cars, creating a congested traffic situation. The cars are parked in various positions along the highway, some closer to the center and others closer to the edges. There are also several trucks visible in the scene, adding to the overall traffic volume. The highway is illuminated by streetlights, providing a clear view of the traffic. The scene is likely a representation of a busy urban highway during rush hour, with a significant amount of traffic moving along the road.

**Figure 8.** RCD effectively reduces visual hallucinations in detailed image descriptions. DoLa's response is omitted when it is identical to the greedy baseline. Correct and hallucinatory contents are highlighted in green and red, respectively. RCD: Retrieval contrastive decoding.

Qualitative samples from our proposed image captioning benchmark are exhibited in Figure 10. RCD enhances the accuracy of image descriptions, mitigating various types of visual hallucinations, such as incorrect physical

**Table 4. Quantitative results on our proposed traffic scenario comprehending test set**

| Model | Decoding | Image captioning | | | | VQA |
|---|---|---|---|---|---|---|
| | | Bleu@4 ↑ | METEOR ↑ | CIDEr ↑ | SPICE ↑ | Accuracy ↑ |
| LLaVA-Next | greedy | 11.4 | 15.7 | 25.0 | 10.8 | 24.7 |
| | +*DoLa* | 11.6 | 15.4 | 24.7 | 10.8 | 25.0 |
| | +*VCD* | 12.1 | 16.6 | 27.4 | 12.0 | 24.6 |
| | +*Ours* | **12.3** | **16.7** | **28.0** | **12.0** | **27.1** |
| LLaVA-1.5 | greedy | **16.5** | 18.0 | 36.7 | 12.6 | 23.2 |
| | +*DoLa* | 16.4 | 18.0 | 36.6 | 12.6 | 23.9 |
| | +*VCD* | 15.9 | 18.2 | 37.5 | 12.6 | 23.9 |
| | +*Ours* | 15.9 | **18.2** | **38.2** | **12.8** | **24.4** |
| InstructBLIP | greedy | 12.2 | 15.5 | 24.8 | 11.1 | 21.3 |
| | +*DoLa* | 12.1 | 15.5 | 24.8 | 11.1 | 21.1 |
| | +*VCD* | 12.6 | **15.6** | 24.9 | **11.3** | 21.3 |
| | +*Ours* | **12.7** | 15.5 | **25.3** | 11.2 | **21.5** |

Our proposed method RCD boosts the image captioning and VQA performance for three MLLMs, outperforming existing methods. Bold font indicates that the method in the corresponding row achieves the highest performance for the corresponding MLLM on the given task. RCD: Retrieval contrastive decoding; MLLMs: multi-modal large language models.

**Table 5. Ablation study on our proposed image captioning benchmark**

| Exp. | img. ret. | diffuse img. | comp. | adapt. | num. refs | Bleu@4 ↑ | METEOR ↑ | CIDEr ↑ | SPICE ↑ |
|---|---|---|---|---|---|---|---|---|---|
| 1 | | | | | n/a | 11.4 | 15.7 | 25.0 | 10.8 |
| 2 | ✓ | ✓ | ✓ | ✓ | 2 | 12.3 | 16.7 | 28.0 | 12.0 |
| 3 | rand. | ✓ | ✓ | ✓ | 2 | $11.9^{\pm0.2}_{(-0.4)}$ | $16.0^{\pm0.1}_{(-0.7)}$ | $26.2^{\pm0.6}_{(-1.8)}$ | $11.9^{\pm0.1}_{(-0.1)}$ |
| 4 | ✓ | | ✓ | ✓ | 2 | 12.5(+0.2) | 15.7(-1.0) | 27.1(-0.9) | 11.3(-0.7) |
| 5 | ✓ | ✓ | add. | ✓ | 2 | 9.6(-2.7) | 14.7(-2.0) | 20.3(-7.7) | 9.9(-2.1) |
| 6 | ✓ | ✓ | ✓ | | 2 | 12.4(+0.1) | 16.3(-0.4) | 26.9(-1.1) | 11.7(-0.3) |
| 7 | ✓ | ✓ | ✓ | ✓ | 1 | 12.6(+0.3) | 16.5(-0.2) | 27.7(-0.3) | 11.9(-0.1) |
| 8 | ✓ | ✓ | ✓ | ✓ | 4 | 13.1(+0.8) | 16.5(-0.2) | 27.9(-0.1) | 11.9(-0.1) |
| 9 | ✓ | ✓ | ✓ | ✓ | 8 | 13.0(+0.7) | 16.4(-0.3) | 27.8(-0.2) | 11.9(-0.1) |

Each component of our proposed RCD positively influences the overall performance. All exper-iments utilize LLaVA-Next-8B as the MLLM (Exp.1). The results for RCD are detailed in Exp.2. The performance differences relative to Exp.2 are shown for Exp.3 through Exp.9.

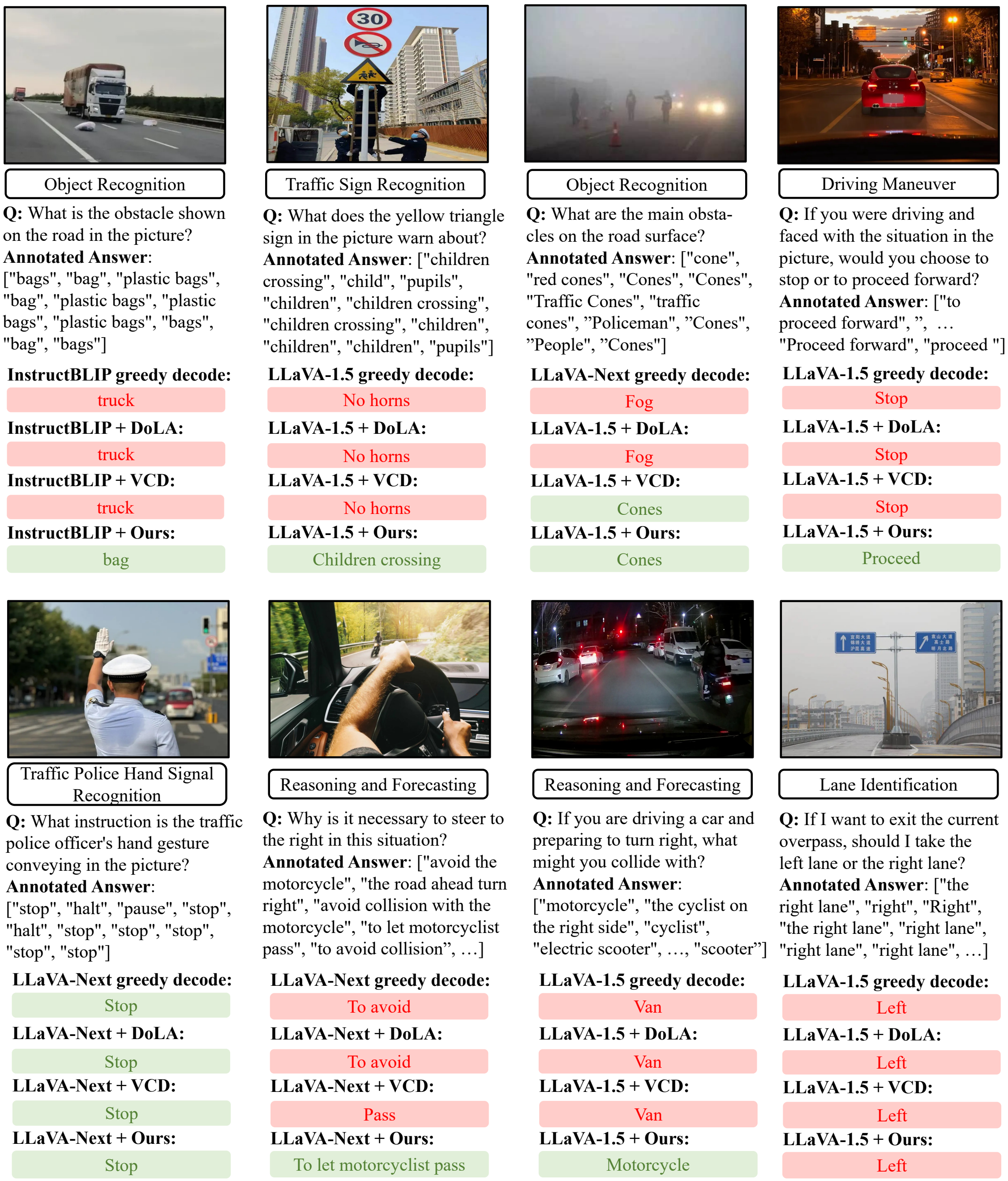
states (e.g., an upside-down car), incorrect object categories (e.g., mistaking a pedestrian sign for a real person), and incorrect spatial relationships (e.g., perceiving a truck coming from the other side). Additionally, RCD improves the specificity of image descriptions, such as identifying three cars involved in a collision on a wet road and a police officer standing in the middle of the street. These examples provide clear evidence of the effectiveness of RCD.

## 5.4. Ablation studies

We assess the effectiveness of each component in our proposed method RCD by systematically ablating them and evaluating the resulting performance variations on the image captioning task of our proposed benchmark. Results are presented in Table 5. Conclusions are summarized as follows:

### *5.4.1 Similar images are better than random images*

We first ablate the image retrieval process (abbreviated img. ret.), opting to randomly sample visual references from the database instead. The average results and standard deviations from five separate trials (each with different random seeds) are presented in Exp.3. The results suggest that random images can positively influence performance (+1.2 CIDEr scores and +1.1 SPICE scores compared to Exp.1), underscoring the resilience of RCD in scenarios lacking similar images. Furthermore, when similar images are available (Exp.2), RCD leads to more substantial improvements across all metrics (e.g., CIDEr +3.0 and SPICE +1.2). Note that the noised



Object Recognition

**Q:** What is the obstacle shown on the road in the picture?
**Annotated Answer**: ["bags", "bag", "plastic bags", "bag", "plastic bags", "plastic bags", "plastic bags", "bags", "bag", "bags"]

**InstructBLIP greedy decode:** 🟥 truck
**InstructBLIP + DoLA:** 🟥 truck
**InstructBLIP + VCD:** 🟥 truck
**InstructBLIP + Ours:** 🟩 bag

Traffic Sign Recognition

**Q:** What does the yellow triangle sign in the picture warn about?
**Annotated Answer**: ["children crossing", "child", "pupils", "children", "children crossing", "children crossing", "children", "children", "children", "pupils"]

**LLaVA-1.5 greedy decode:** 🟥 No horns
**LLaVA-1.5 + DoLA:** 🟥 No horns
**LLaVA-1.5 + VCD:** 🟥 No horns
**LLaVA-1.5 + Ours:** 🟩 Children crossing

Object Recognition

**Q:** What are the main obstacles on the road surface?
**Annotated Answer**: ["cone", "red cones", "Cones", "Cones", "Traffic Cones", "traffic cones", ”Policeman", ”Cones", ”People", ”Cones"]

**LLaVA-Next greedy decode:** 🟥 Fog
**LLaVA-1.5 + DoLA:** 🟥 Fog
**LLaVA-1.5 + VCD:** 🟩 Cones
**LLaVA-1.5 + Ours:** 🟩 Cones

Driving Maneuver

**Q:** If you were driving and faced with the situation in the picture, would you choose to stop or to proceed forward?
**Annotated Answer**: ["to proceed forward", ", … "Proceed forward", "proceed "]

**LLaVA-1.5 greedy decode:** 🟥 Stop
**LLaVA-1.5 + DoLA:** 🟥 Stop
**LLaVA-1.5 + VCD:** 🟥 Stop
**LLaVA-1.5 + Ours:** 🟩 Proceed

Traffic Police Hand Signal Recognition

**Q:** What instruction is the traffic police officer's hand gesture conveying in the picture?
**Annotated Answer**: ["stop", "halt", "pause", "stop", "halt", "stop", "stop", "stop", "stop", "stop"]

**LLaVA-Next greedy decode:** 🟩 Stop
**LLaVA-Next + DoLA:** 🟩 Stop
**LLaVA-Next + VCD:** 🟩 Stop
**LLaVA-Next + Ours:** 🟩 Stop

Reasoning and Forecasting

**Q:** Why is it necessary to steer to the right in this situation?
**Annotated Answer**: ["avoid the motorcycle", "the road ahead turn right", "avoid collision with the motorcycle", "to let motorcyclist pass", "to avoid collision”, …]

**LLaVA-Next greedy decode:** 🟥 To avoid
**LLaVA-Next + DoLA:** 🟥 To avoid
**LLaVA-Next + VCD:** 🟥 Pass
**LLaVA-Next + Ours:** 🟩 To let motorcyclist pass

Reasoning and Forecasting

**Q:** If you are driving a car and preparing to turn right, what might you collide with?
**Annotated Answer**: ["motorcycle", "the cyclist on the right side", "cyclist", "electric scooter", …, "scooter”]

**LLaVA-1.5 greedy decode:** 🟥 Van
**LLaVA-1.5 + DoLA:** 🟥 Van
**LLaVA-1.5 + VCD:** 🟥 Van
**LLaVA-1.5 + Ours:** 🟩 Motorcycle

Lane Identification

**Q:** If I want to exit the current overpass, should I take the left lane or the right lane?
**Annotated Answer**: ["the right lane", "right", "Right", "the right lane", "right lane", "right lane", "right lane", …]

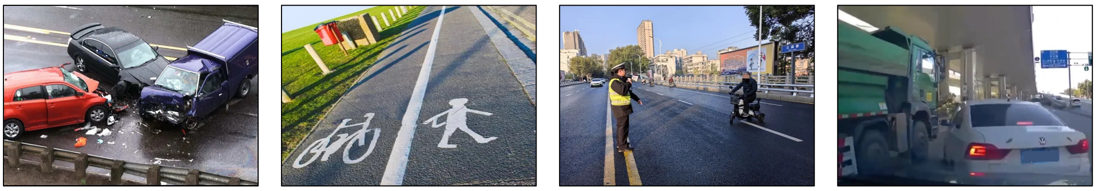
**LLaVA-1.5 greedy decode:** 🟥 Left
**LLaVA-1.5 + DoLA:** 🟥 Left
**LLaVA-1.5 + VCD:** 🟥 Left
**LLaVA-1.5 + Ours:** 🟥 Left

**Figure 9.** Qualitative results on our proposed VQA benchmark. Our method enhances the accuracy of answers from three MLLMs across nearly all question categories. Human-annotated answers are provided for each test sample. Incorrect and correct answers are highlighted in red green, respectively. VQA: Visual Question Answering.

image is retained because it aids in mitigating erroneous language priors, and the CIDEr score dropped by 0.9 in Exp.4 after it is ablated.

#### *5.4.2 Visual concepts comparison is better than combination*

We modify the visual concept comparison process (abbreviated comp.), transitioning from logit subtraction in Equation (6) into addition. The additive paradigm emphasizes the shared visual cues across similar images. Exp.5 demonstrates significant declines in all metrics (e.g.,-7.7 CIDEr score and -2.1 SPICE score). These results support the superiority of contrasting visual concepts between similar images over combining them.



| | | | |
|---|---|---|---|
| **LLaVA-1.5 greedy decode:** A red car is upside down on the road and is surrounded by other cars. | **LLaVA-1.5 greedy decode:** A bike lane is painted on the road with a man walking on the other side. | **LLaVA-Next greedy decode:** A man on a motorcycle talking to a police officer. | **InstructBLIP greedy decode:** a car is driving under a bridge and a truck is coming from the other side |
| **LLaVA-1.5 + DoLA:** A red car is upside down on the road and is surrounded by other cars. | **LLaVA-1.5 + DoLA:** A bike lane is painted on the road with a person walking on the other side. | **LLaVA-Next + DoLA:** A man on a motorcycle talking to a police officer. | **InstructBLIP + DoLA:** a car is driving under a bridge and a truck is coming. |
| **LLaVA-1.5 + VCD:** A red car is upside down on the road and is next to a black car. | **LLaVA-1.5 + VCD:** A bike lane is painted on the road with a man walking on the other side. | **LLaVA-Next + VCD:** A policeman directing traffic on a city street. | **InstructBLIP + VCD:** a truck is driving down the street next to a car. |
| **LLaVA-1.5 + Ours:** Three cars are involved in a crash on a wet road. | **LLaVA-1.5 + Ours:** A bike lane is painted on the road with a pedestrian symbol walking. | **LLaVA-Next + Ours:** A police officer standing in the middle of a street talking to a scooter rider. | **InstructBLIP + Ours:** a truck is driving down the street next to a car. |

**Figure 10.** Qualitative results on our proposed image captioning benchmark. Our method RCD improves the accuracy of image descriptions and reduces various types of visual hallucinations. Additionally, RCD enhances the specificity of the descriptions. Incorrect content is highlighted in red, while correct content is highlighted in green. RCD: Retrieval contrastive decoding.

#### *5.4.3 Other ablations*

Exp.6 demonstrates that the adaptive scaling strategy (abbreviated as adapt.) can enhance almost all metrics. Additionally, increasing the number of reference images yields improvements on the BLEU@4 metric (from 12.3 to 13.0). However, it also leads to slight decreases on the CIDEr score (from 28.0 to 27.8) in Exp.7-9. Therefore, RCD defaults to using two visual references.

### 5.5. Further experimental analysis

#### *5.5.1. Analysis of various text decoding baselines*

In the experiments on public benchmarks, greedy search is used as the baseline strategy. This approach selects only the token candidate with the highest confidence score, thereby excluding any randomness in text decoding. We further investigate the integration of RCD with various token sampling baselines, including multi-nominal sampling, top-$k$ sampling, and top-$p$ sampling. The experimental results are presented in Table 6. Notably, RCD integrates well with different baseline sampling strategies. However, these token sampling strategies exhibit significantly weaker performance compared to greedy decoding.

#### *5.5.2. Analysis of similarity metrics*

RCD employs cosine similarity as the similarity measure for reference information retrieval. Table 7 provides a quantitative analysis on our proposed traffic scene image captioning test set comparing cosine similarity with the L2 distance metric. These results indicate that the performance based on the Euclidean distance metric is marginally inferior to that of the cosine similarity metric, yet it consistently surpasses the baseline performance of the greedy decoding approach.

#### *5.5.3. Analysis of the reference database's size*

RCD uses a reference database containing 113k samples from the COCO Caption dataset. Table 8 demonstrates that incorporating the entire Visual Genome dataset[63] into the reference database doubles the size of the database but leads to a decline in the evaluation metrics of the LLaVA-1.5 model. For instance, the CIDEr

**Table 6. Image captioning performance with various token sampling baselines on the WHOOPS benchmark**

| | Method | B4 ↑ | M ↑ | C ↑ | S ↑ |
|---|---|---|---|---|---|
| LLaVA-1.5 | greedy | 19.7 | 25.6 | 67.9 | 17.3 |
| | +*RCD* | **20.0** | **26.3** | **75.5** | **17.8** |
| | multi-nominal sampling | 6.9 | 19.0 | 31.4 | 12.0 |
| | +*RCD* | 8.8 | 20.3 | 40.6 | 13.7 |
| | nucleus sampling | 9.6 | 20.5 | 39.4 | 13.3 |
| | +*RCD* | 10.7 | 21.8 | 46.8 | 14.0 |
| | Top-k sampling | 8.0 | 19.3 | 33.6 | 12.3 |
| | +*RCD* | 8.8 | 20.3 | 40.6 | 13.7 |

RCD boosts model performance on various text decoding baselines. For nu-cleus (Top-$p$) sampling, $p$ is set at 0.9. For Top-$k$ sampling, $k$ is set at 50, following common practice. The random seeds are fixed at 42 in all exper-iments. Bold font indicates the highest performance. RCD: Retrieval con-trastive decoding.

**Table 7. Quantitative results on our proposed traffic scenario description test set**

| Model | Method | B4 ↑ | M ↑ | C ↑ | S ↑ |
|---|---|---|---|---|---|
| LLaVA-Next | greedy | 11.4 | 15.7 | 25.0 | 10.8 |
| | RCD-L2 | 12.1 | 16.3 | 27.0 | 11.8 |
| | RCD-Cos | **12.3** | **16.7** | **28.0** | **12.0** |

L2 denotes using Euclidean distance as similarity metric for RCD. Cos de-notes cosine similarity. Bold font indicates the highest performance. RCD: Retrieval contrastive decoding.

**Table 8. Experiment on WHOOPS image captioning benchmark with enlarged reference database for LLaVA-1.5**

| Model | Method | # Ref. | B4 ↑ | M ↑ | C ↑ | S ↑ |
|---|---|---|---|---|---|---|
| LLaVA-1.5 | greedy | - | 19.7 | 25.6 | 67.9 | 17.3 |
| | COCO | 113.2k | **20.0** | **26.3** | **75.5** | **17.8** |
| | COCO+VG | 220.6k | 20.0 | 26.3 | 74.6 | 17.7 |

"# Ref." denotes the number of samples in the reference database. Bold font indicates the highest performance.

score slightly decreases by 0.9 compared to using the COCO Caption data. This decline is likely attributable to the increased inaccuracy in image retrieval results, as the reference database becomes significantly more expansive. As a result, the retrieval module's robustness may require further optimization to effectively manage larger-scale reference databases.

### *5.5.4. Analysis of inference efficiency*

RCD requires the independent prediction of an additional $k+1$ confidence score distribution, which introduces linear computational complexity. As shown in Table 9, the naive sequential prediction of these distributions (implemented using a *for* loop, denoted as RCD-Naive) increases the MLLM's inference time by approximately a factor of $k$. To mitigate this delay, the $k+1$ input sequences can be stacked for batch processing. This optimization (RCD-Optimized) reduces the $k$-fold increase in inference time to approximately $0.43k$-fold. Moreover, as demonstrated in Table 9, both implementations of RCD result in only a negligible increase in GPU memory (less than 15 MB) when $k = 2$.

## 6. DISCUSSION

### 6.1. Key contributions

Through extensive experiments, we find that MLLMs' visual branch often equally supports both erroneous and accurate token candidates. This indicates that MLLMs are not blind amid hallucinations, and suggests

**Table 9. Efficiency analysis on our proposed traffic scene VQA benchmark**

| Model | Method | $k$ | Time (s) | Memory (MB) |
|---|---|---|---|---|
| LLaVA-Next | greedy | - | $0.7684_{\pm 0.0065}$ | 19569 |
| | RCD-Naive | 2 | $1.6498_{\pm 0.0034}$ | 19581 |
| | RCD-Optimized | 2 | $1.1499_{\pm 0.0012}$ | 19579 |

"Time" denotes the wall clock time for a single test sample, which is averaged on 200 samples from three separate runs. RCD: Retrieval contrastive decoding.

new perspectives for visual hallucination mitigation by eliciting accurate content from non-blind MLLMs. We further provide quantitative and qualitative evidence to reveal that analogous hallucinations can occur among similar images, and this phenomenon can be exploited to distinguish accurate visual content. We hope that our findings can inspire subsequent research on the hallucination issue of MLLMs.

Based on our analysis on the characteristics of visual hallucinations, we introduce a novel method named RCD to mitigate visual hallucinations. RCD retrieves relevant images to serve as references for MLLMs during visual recognition. This process is analogous to the memory recall mechanism in the human brain. Next, RCD distinguishes accurate visual cues by contrasting the visual cues between the test image and the visual references. Our proposed method is a plug-and-play module that can be seamlessly integrated into MLLMs to reduce hallucinations originating from both visual and textual branches of MLLMs. Quantitative and qualitative results demonstrate that RCD improves the accuracy of visual recognition and enhances the reasoning capabilities of MLLMs.

### 6.2. Limitations and future work

Our proposed method RCD is based on the premise that images with similar semantics and appearance are likely to induce analogous visual hallucinations. We provide both quantitative and qualitative evidence in Section 3.2 to support this hypothesis, specifically in relation to the widely used LLaVA model family (LLaVA-1.5 and LLaVA-Next). Additionally, we also observe significant performance improvements on two hallucination benchmarks when RCD is applied to InstructBLIP. In future work, we aim to investigate whether this phenomenon is dependent on the model or if it is model-independent.

A reference database consisting of around 113,000 samples from the COCO Caption dataset is utilized in all experiments. These images encompass a rich variety of scenes and objects, including real-world traffic scenarios. The diversity of visual references enables RCD to enhance model performance across multiple evaluation benchmarks. However, experimental results demonstrate that there is still room for improvement in the retrieval module's ability to handle larger databases. Additionally, improvements are needed in the current state-of-the-art MLLMs on our proposed traffic scenario comprehension benchmark. We plan to conduct a more detailed analysis of the types of failure cases, and optimize the performance of RCD.

## 7. CONCLUSIONS

We address the issue of visual hallucinations in MLLMs. Through comprehensive analysis, we find that even in the presence of visual hallucinations, MLLMs can still recognize accurate visual cues. Furthermore, we demonstrate that analogous visual hallucinations induced by similar images can be exploited to mitigate visual hallucinations. Building on this insight, we introduce RCD, a training-free method that effectively mitigates visual hallucinations in MLLMs. RCD retrieves similar images from a database to serve as references for MLLMs, and discerns accurate visual content through confidence score comparisons. This approach corrects erroneous content that is mistakenly supported by both the visual and textual branches of MLLMs. Experiments conducted on two publicly available hallucination benchmarks demonstrate the superiority of our proposed method. RCD leads to significant performance improvements in three leading MLLMs: InstructBLIP, LLaVA-

1.5 and LLaVA-Next. Additionally, RCD outperforms existing advanced decoding strategies. We also curate a benchmark focused on understanding real-world traffic scenarios, which includes challenging questions designed to evaluate the effectiveness of RCD in assisting MLLMs in practical applications. Both quantitative and qualitative results demonstrate that RCD significantly enhances the ability of three MLLMs to accurately comprehend real-world traffic scenes. This improved visual comprehension forms the foundation for developing reliable embodied AI systems that employ MLLMs in real-world applications.

## DECLARATIONS

### Acknowledgments

We thank the Editor-in-Chief and all reviewers for their comments. We thank the volunteers for their excellent work in data annotation. We thank Professor Changjun Jiang and Professor Alois Knoll for their discussions and feedback on this reseach.

### Authors' contributions

Conceived the main idea: Yang, D.; Chen, G.
Supervised this work: Chen, G.
Developed the code for the core algorithm: Yang, D.
Designed and performed the analysis for visual hallucinations: Yang, D.; Cao, B.
Collected and processed the traffic scenario comprehension benchmark: Yang, D.
Wrote the initial version of the manuscript: Yang, D.; Cao, B.
Revised the manuscript: Yang, D.; Cao, B.; Qu, S.; Lu, F.; Gu, S.; Chen, G.

### Availability of data and materials

The MME and POPE datasets used for evaluation in this study were sourced from their official repository (https://github.com/RUCAIBox/POPE) and (https://github.com/BradyFU/Awesome-Multimodal-Large-Language-Models/tree/Evaluation), respectively. The data for our proposed benchmark will be made available upon request; please contact the corresponding author via email if needed.

### Financial support and sponsorship

This work was supported by the National Key Research and Development Program of China (No. 2024YFE0211000), in part by the National Natural Science Foundation of China (No. 62372329), in part by Shanghai Scientific Innovation Foundation (No. 23DZ1203400), in part by Tongji-Qomolo Autonomous Driving Commercial Vehicle Joint Lab Project, and in part by Xiaomi Young Talents Program.

### Conflicts of interest

Chen, G. is a member of the Editorial Board of the journal *Intelligence & Robotics*. Chen, G. was not involved in any steps of editorial processing, notably including reviewer selection, manuscript handling, or decision-making. The other authors declare that there are no conflicts of interest.

### Ethical approval and consent to participate

Not applicable.

### Consent for publication

Not applicable.

### Copyright

## REFERENCES

1. Tong S.; Brown E.; Wu P.; et al. Cambrian-1: a fully open, vision-centric exploration of multimodal LLMs. *arXiv* **2024**, arXiv:2406.16860. Available online: https://doi.org/10.48550/arXiv.2406.16860. (accessed on 12 Mar 2025)
2. Liu, H.; Li, C.; Li, Y.; Lee, Y. J. Improved baselines with visual instruction tuning. *arXiv* **2023**, arXiv:2310.03744. Available online: https://doi.org/10.48550/arXiv.2310.03744. (accessed on 12 Mar 2025)
3. Dai, W.; Li, J.; Li, D.; et al. InstructBLIP: towards general-purpose vision-language models with instruction tuning. *arXiv* **2023**, arXiv:2305.06500. Available online: https://doi.org/10.48550/arXiv.2305.06500. (accessed on 12 Mar 2025)
4. Zhu, D.; Chen, J.; Shen, X.; Li, X.; Elhoseiny, M. MiniGPT-4: enhancing vision-language understanding with advanced large language models. *arXiv* **2023**, arXiv:2304.10592. Available online: https://doi.org/10.48550/arXiv.2304.10592. (accessed on 12 Mar 2025)
5. Lin, Z.; Liu, C.; Zhang, R.; et al. Sphinx: the joint mixing of weights, tasks, and visual embeddings for multi-modal large language models. *arXiv* **2023**, arXiv:2311.07575. Available online: https://doi.org/10.48550/arXiv.2311.07575. (accessed on 12 Mar 2025)
6. Li, J.; Li, D.; Savarese, S.; Hoi, S. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv* **2023**, arXiv:2301.12597. Available online: https://doi.org/10.48550/arXiv.2301.12597. (accessed on 12 Mar 2025)
7. You, H.; Zhang, H.; Gan, Z.; et al. Ferret: refer and ground anything anywhere at any granularity. *arXiv* **2023**, arXiv:2310.07704. Available online: https://doi.org/10.48550/arXiv.2310.07704. (accessed on 12 Mar 2025)
8. Yuan, Y.; Li, W.; Liu, J.; et al. Osprey: pixel understanding with visual instruction tuning. *arXiv* **2023**, arXiv:2312.10032. Available online: https://doi.org/10.48550/arXiv.2312.10032. (accessed on 12 Mar 2025)
9. Driess, D.; Xia, F.; Sajjadi, M. S.; et al. PaLM-E: an embodied multimodal language model. *arXiv* **2023**, arXiv:2303.03378. Available online: https://doi.org/10.48550/arXiv.2303.03378. (accessed on 12 Mar 2025)
10. Xu, Z.; Zhang, Y.; Xie, E.; et al. DriveGPT4: interpretable end-to-end autonomous driving via large language model. *IEEE. Robot. Autom. Lett.* **2024**, *9*, 8186-93. DOI
11. Cui, C.; Ma, Y.; Cao, X.; et al. A survey on multimodal large language models for autonomous driving. *arXiv* **2023**, arXiv:2311.12320. Available online: https://arxiv.org/abs/2311.12320. (accessed on 12 Mar 2025)
12. Liu H.; Xue W.; Chen Y.; et al. A survey on hallucination in large vision-language models. *arXiv* **2024**, arXiv:2402.00253. Available online: https://doi.org/10.48550/arXiv.2402.00253. (accessed on 12 Mar 2025)
13. Huang, Q.; Dong, X.; Zhang, P.; et al. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. *arXiv* **2023**, arXiv:2311.17911. Available online: https://doi.org/10.48550/arXiv.2311.17911. (accessed on 12 Mar 2025)
14. Jiang, C.; Xu, H.; Dong, M.; et al. Hallucination augmented contrastive learning for multimodal large language model. *arXiv* **2023**, arXiv:2312.06968. Available online: https://doi.org/10.48550/arXiv.2312.06968. (accessed on 12 Mar 2025)
15. Leng, S.; Zhang, H.; Chen, G.; et al. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2024. pp. 13872–82. DOI
16. Yin, S.; Fu, C.; Zhao, S.; et al. Woodpecker: hallucination correction for multimodal large language models. *Sci. China. Inf. Sci.* **2024**, *67*, 220105. DOI
17. Zhou, Y.; Cui, C.; Yoon, J.; et al. Analyzing and mitigating object hallucination in large vision-language models. *arXiv* **2023**, arXiv:2310.00754. Available online: https://doi.org/10.48550/arXiv.2310.00754. (accessed on 12 Mar 2025)
18. Tong, S.; Liu, Z.; Zhai, Y.; Ma, Y.; LeCun, Y.; Xie, S. Eyes wide shut? Exploring the visual shortcomings of multimodal LLMs. In: *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, Seattle, USA. Jun 16-22, 2024. IEEE, 2024; pp. 9568–78. DOI
19. Liu, F.; Lin, K.; Li, L.; Wang, J.; Yacoob, Y.; Wang L. Mitigating hallucination in large multi-modal models via robust instruction tuning. *arXiv* **2023**, arXiv:2306.14565. Available online: https://doi.org/10.48550/arXiv.2306.14565. (accessed on 12 Mar 2025)
20. Lee, S.; Park, S. H.; Jo, Y.; Seo, M. Volcano: mitigating multimodal hallucination through self-feedback guided revision. *arXiv* **2023**, arXiv:2311.07362. Available online: https://doi.org/10.48550/arXiv.2311.07362. (accessed on 12 Mar 2025)
21. Favero, A.; Zancato, L.; Trager, M.; et al. Multi-modal hallucination control by visual information grounding. In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*; 2024. pp. 14303–12. DOI
22. Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; Parikh, D. Making the V in VQA matter: elevating the role of image understanding in visual question answering. In: *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Honolulu, USA. Jul 21-26, 2017. IEEE, 2017; pp. 6904–13. DOI
23. Li, B.; Zhang, K.; Zhang, H.; et al. LLaVA-NeXT: stronger LLMs supercharge multimodal capabilities in the wild. 2024. https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/. (accessed 2025-03-12)
24. Fu C.; Chen P.; Shen Y.; et al. MME: a comprehensive evaluation benchmark for multimodal large language models. *arXiv* **2023**, arXiv:2306.13394. Available online: https://doi.org/10.48550/arXiv.2306.13394. (accessed on 12 Mar 2025)
25. Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, W. X.; Wen, J. R. Evaluating object hallucination in large vision-language models. *arXiv* **2023**, arXiv:2305.10355. Available online: https://doi.org/10.48550/arXiv.2305.10355. (accessed on 12 Mar 2025)
26. Bitton-Guetta, N.;, Bitton, Y.; Hessel, J.; et al. Breaking common sense: Whoops! A vision-and-language benchmark of synthetic and compositional images. In: *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, Paris, France. Oct 01-06, 2023. IEEE, 2023; pp. 2616–27. DOI
27. Liu, H.; Li, C.; Wu, Q.; Lee, Y. J. Visual instruction tuning. *arXiv* **2024**, arXiv:2304.08485. Available online: https://doi.org/10.48550/arXiv.2304.08485. (accessed on 12 Mar 2025)
28. Jiang, K.; Wang, Z.; Yi, P.; Lu, T.; Jiang, J.; Xiong, Z. Dual-path deep fusion network for face image hallucination. *IEEE. Trans. Neur.*

*Net. Learn. Syst.* **2020**, *33*, 378–91. DOI

29. Rohrbach, A.; Hendricks, L. A.; Burns, K.; Darrell, T.; Saenko, K. Object hallucination in image gaptioning. *arXiv* **2018**, arXiv:1809.02156. Available online: https://doi.org/10.48550/arXiv.1809.02156. (accessed on 12 Mar 2025)
30. Wu M.; Ji J.; Huang O.; et al. Evaluating and analyzing relationship hallucinations in large vision-language models. *arXiv* **2024**, arXiv:2406.16449. Available online: https://doi.org/10.48550/arXiv.2406.16449. (accessed on 12 Mar 2025)
31. Sun Y.; Zhang Z.; Wu H.; et al. Explore the hallucination on low-level perception for MLLMs. *arXiv* **2024**, arXiv:2409.09748. Available online: https://doi.org/10.48550/arXiv.2409.09748. (accessed on 12 Mar 2025)
32. Shi, W.; Han, X.; Lewis, M.; Tsvetkov, Y.; Zettlemoyer, L.; Yih, S. W. Trusting your evidence: hallucinate less with context-aware decoding. *arXiv* **2023**, arXiv:2305.14739. Available online: https://doi.org/10.48550/arXiv.2305.14739. (accessed on 12 Mar 2025)
33. Zhang, M.; Press, O.; Merrill, W.; Liu A.; Smith, N. A. How language model hallucinations can snowball. *arXiv* **2023**, arXiv:2305.13534. Available online: https://doi.org/10.48550/arXiv.2305.13534. (accessed on 12 Mar 2025)
34. Yu, T.; Zhang, H.; Yao, Y.; et al. RLAIF-V: aligning MLLMs through open-source AI feedback for super GPT-4V trustworthiness. 2024. https://openreview.net/forum?id=iRa9PK0opY. (accessed on 2025-03-12)
35. Xie, Y.; Li, G.; Xu, X.; Kan, M. Y. V-DPO: mitigating hallucination in large vision language models via vision-guided direct preference optimization. *arXiv* **2024**, arXiv:2411.02712. Available online: https://doi.org/10.48550/arXiv.2411.02712. (accessed on 12 Mar 2025)
36. Ouali, Y.; Bulat, A.; Martinez, B.; Tzimiropoulos, G. CLIP-DPO: vision-language models as a source of preference for fixing hallucinations in LVLMs. *arXiv* **2024**, arXiv:2408.10433. Available online: https://doi.org/10.48550/arXiv.2408.10433. (accessed on 12 Mar 2025)
37. Rafailov, R.; Sharma, A.; Mitchell, E.; Ermon, S.; Manning, C. D.; Finn, C. Direct preference optimization: Your language model is secretly a reward model. *arXiv* **2023**, arXiv:2305.18290. Available online: https://doi.org/10.48550/arXiv.2305.18290. (accessed on 12 Mar 2025)
38. Chen, Z.; Zhao, Z.; Luo, H.; Yao, H.; Li, B.; Zhou, J. HALC: object hallucination reduction via adaptive focal-contrast decoding. In: *Forty-first International Conference on Machine Learning*; 2024. https://openreview.net/forum?id=EYvEVbfoDp. (accessed 2025-03-12)
39. Liang, X.; Yu, J.; Mu, L.; et al. Mitigating hallucination in visual-language models via re-balancing contrastive decoding. *arXiv* **2024**, arXiv:2409.06485. Available online: https://doi.org/10.48550/arXiv.2409.06485. (accessed on 12 Mar 2025)
40. Zhao, Z. H.; Wallace, E.; Feng S.; Klein, D.; Singh, S. Calibrate before use: improving few-shot performance of language models. *arXiv* **2021**, arXiv:2102.09690. Available online: https://doi.org/10.48550/arXiv.2102.09690. (accessed on 12 Mar 2025)
41. Li, X. L.; Holtzman, A.; Fried, D.; et al. Contrastive decoding: open-ended text generation as optimization. *arXiv* **2022**, arXiv:2210.15097. Available online: https://doi.org/10.48550/arXiv.2210.15097. (accessed on 12 Mar 2025)
42. Zheng, L.; Chiang, W. L.; Sheng Y.; et al. Judging LLM-as-a-judge with MT-bench and chatbot arena. *arXiv* **2023**, arXiv:2306.05685. Available online: https://doi.org/10.48550/arXiv.2306.05685. (accessed on 12 Mar 2025)
43. Touvron, H.; Martin, L.; Stone, K.; et al. Llama 2: open foundation and fine-tuned chat models. *arXiv* **2023**, arXiv:2307.09288. Available online: https://doi.org/10.48550/arXiv.2307.09288. (accessed on 12 Mar 2025)
44. Vaswani, A.; Shazeer, N.; Parmar, N.; et al. Attention is all you need. *arXiv* **2017**, arXiv:1706.03762. Available online: https://doi.org/10.48550/arXiv.1706.03762. (accessed on 12 Mar 2025)
45. Lin, B. Y.; Ravichander, A.; Lu, X.; et al. The unlocking spell on base LLMs: rethinking alignment via in-context learning. In: *The Twelfth International Conference on Learning Representations*; 2024. https://openreview.net/forum?id=wxJ0eXwwda. (accessed on 2025-03-12)
46. Kuznetsova, A.; Rom, H.; Alldrin, N.; et al. The open images dataset V4: unified image classification, object detection, and visual relationship detection at scale. *Int. J. Comput. Vis.* **2020**, *128*, 1956–81. DOI
47. Lin, T. Y.; Maire, M.; Belongie, S.; et al. Microsoft COCO: common objects in context. In: *Computer Vision–ECCV 2014: 13th European Conference*, Zurich, Switzerland. Sep 6-12, 2014. Springer, 2014; pp. 740–55. DOI
48. Ho, J.; Jain, A. N.; Abbeel, P. Denoising diffusion probabilistic models. *Adv. Neural Inform. Process. Syst.* **2020**, *33*, 6840–51. https://proceedings.neurips.cc/paper/2020/hash/4c5bcfec8584af0d967f1ab10179ca4b-Abstract.html. (accessed on 2025-03-12)
49. Xiao, Y.; Yuan, Q.; Jiang, K.; He, J.; Jin, X.; Zhang, L. EDiffSR: an efficient diffusion probabilistic model for remote sensing image super-resolution. *IEEE. Trans. Geosci. Remote. Sens.* **2023**, *62*, 1–14. DOI
50. Yan, P.; Li, M.; Zhang, J.; Li, G.; Jiang, Y.; Luo, H. Cold SegDiffusion: a novel diffusion model for medical image segmentation. *Knowl. Based. Syst.* **2024**, *301*, 112350. DOI
51. Anciukevčius, T.; Xu, Z. X.; Fisher, M.; et al. Renderdiffusion: Image diffusion for 3D reconstruction, inpainting and generation. *arXiv* **2022**, arXiv:2211.09869. Available online: https://doi.org/10.48550/arXiv.2211.09869. (accessed on 12 Mar 2025)
52. Andrej, K.; Li, F. F. Deep visual-semantic alignments for generating image descriptions. In: *Proceedings of the IEEE conference on computer vision and pattern recognition*; 2015. pp. 3128–37. DOI
53. Radford, A.; Kim, J. W.; Hallacy, C.; et al. Learning transferable visual models from natural language supervision. *arXiv* **2021**, arXiv:2103.00020. Available online: https://doi.org/10.48550/arXiv.2103.00020. (accessed on 12 Mar 2025)
54. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; et al. An image is worth 16×16 words: transformers for image recognition at scale. *arXiv* **2020**, arXiv:2010.11929. Available online: https://doi.org/10.48550/arXiv.2010.11929. (accessed on 12 Mar 2025)
55. Oquab, M.; Darcet, T.; Moutakanni, T.; et al. DINOv2: learning robust visual features without supervision. 2024. https://openreview.net/forum?id=a68SUt6zFt. (accessed on 2025-03-12)
56. Douze, M.; Guzhva, A.; Deng, C.; et al. The Faiss library. *arXiv* **2024**, arXiv:2401.08281. Available online: https://doi.org/10.48550/arXiv.2401.08281. (accessed on 12 Mar 2025)
57. Papineni, K.; Roukos, S.; Ward, T.; Zhu, W. J. BLEU: a method for automatic evaluation of machine translation. In: *Proceedings*

*of the 40th annual meeting of the Association for Computational Linguistics*. 2002. pp. 311–18. https://aclanthology.org/P02-1040.Pdf. (accessed on 2025-03-12)

58. Denkowski, M.; Lavie, A. Meteor universal: language specific translation evaluation for any target language. In: *Proceedings of the Ninth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, 2014; pp. 376–80. DOI
59. Vedantam, R.; Lawrence Zitnick, C.; Parikh, D. CIDEr: consensus-based image description evaluation. In: *2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Boston, USA. Jun 07-12, 2015. IEEE, 2015; pp. 4566–75. DOI
60. Anderson, P.; Fernando, B.; Johnson, M.; Gould, S. SPICE: semantic propositional image caption evaluation. In: *Computer Vision–ECCV 2016: 14th European Conference*, Amsterdam, The Netherlands. Oct 11-14, 2016. Springer, 2016; pp. 382–98. DOI
61. Paszke, A.; Gross, S.; Massa, F.; et al. PyTorch: an imperative style, high-performance deep learning library. *arXiv* **2019**, arXiv:1912.01703. Available online: https://doi.org/10.48550/arXiv.1912.01703. (accessed on 12 Mar 2025)
62. Chuang, Y. S.; Xie, Y.; Luo, H.; Kim, Y.; Glass, J. R.; He, P. DoLa: decoding by contrasting layers improves factuality in large language models. In: *The Twelfth International Conference on Learning Representations*; 2024. https://openreview.net/forum?id=Th6NyL07na. (accessed on 2025-03-12)
63. Krishna, R.; Zhu, Y.; Groth, O.; et al. Visual genome: connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.* **2017**, *123*, 32–73. DOI