# RRHF-V: Ranking Responses to Mitigate Hallucinations in Multimodal Large Language Models with Human Feedback

**Guoqing Chen, Fu Zhang*, Jinghao Lin, Chenglong Lu, Jingwei Cheng**
School of Computer Science and Engineering, Northeastern University, China
Key Laboratory of Intelligent Computing in Medical Image of Ministry of Education
Northeastern University, China
chenguoqing247@gmail.com; {zhangfu,chengjingwei}@mail.neu.edu.cn

## Abstract

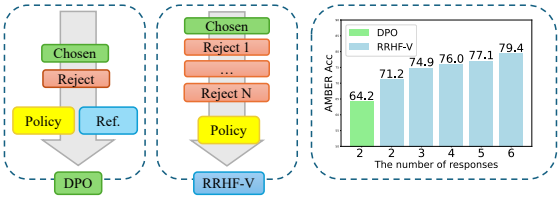
Multimodal large language models (MLLMs) demonstrate strong capabilities in multimodal understanding, reasoning, and interaction but still face the fundamental limitation of hallucinations, where they generate erroneous or fabricated information. To mitigate hallucinations, existing methods annotate pair-responses (one non-hallucination *vs* one hallucination) using manual methods or GPT-4V, and train alignment algorithms to improve the correspondence between images and text. More critically, an image description often involve multiple dimensions (e.g., object attributes, posture, and spatial relationships), making it challenging for the model to comprehensively learn multi-dimensional information from pair-responses.

To this end, in this paper, we propose **RRHF-V**, which is the first using rank-responses (one non-hallucination *vs* multiple ranking hallucinations) to mitigate multimodal hallucinations. Instead of using pair-responses to train the model, RRHF-V expands the number of hallucinatory responses, so that the responses with different scores in a rank-response enable the model to learn rich semantic information across various dimensions of the image. Further, we propose a scene graph-based approach to automatically construct rank-responses in a cost-effective and automatic manner. We also design a novel training objective based on rank loss and margin loss to balance the differences between hallucinatory responses within a rank-response, thereby improving the model's image comprehension. Experiments on two MLLMs of different sizes and four widely used benchmarks demonstrate that RRHF-V is effective in mitigating hallucinations and outperforms the DPO method based on pair-responses.[1]

## 1 Introduction

Large Language Models (LLMs) (OpenAI, 2023; Touvron et al., 2023; Achiam et al., 2023; Jiang et al., 2023; Bai et al., 2023a) represent a significant milestone in the field of natural language processing. They have been further extended to encompass multimodal domains, leading to the emergence of Multimodal Large Language Models (MLLMs) (Achiam et al., 2023; Liu et al., 2024c; Team et al., 2023; Bai et al., 2023b). Despite their impressive capabilities, which enable them to excel in various visual tasks (Zhang et al., 2023; Li et al., 2024; Black et al., 2024) and handle complex content understanding (Lai et al., 2024) or generation (Brooks et al., 2023; Geng et al., 2024), MLLMs face a major challenge known as the "*Hallucination*" problem. Specifically, MLLMs frequently generate incorrect statements in responses to user-provided images and prompts, such as producing irrelevant or nonsensical responses, identifying non-existent colors, inaccuracies in quantities, and incorrect object positions in images. This flaw poses significant risks to the practical application of MLLMs, making them less reliable as assistants.

Various approaches have been proposed to mitigate hallucinations in MLLMs. These methods




(a) Pair-response (b) Rank-response (c) Hallucination Evaluation Results

Figure 1: Figure (a) and (b) illustrate a comparison between DPO (Rafailov et al., 2024) and our RRHF-V. "Chosen" and "Reject" represent non-hallucinatory and hallucinatory responses, respectively. RRHF-V training with our automatically constructed rank-responses does not require a reference (Ref.) model, and the number of hallucinatory responses extends from 1 to *N*. In Figure (c), our exploratory experiments on the hallucination evaluation metric AMBER Acc (Wang et al., 2023) demonstrate the effectiveness of using rank-responses.

[1]Code: https://github.com/chengq1001/RRHF-V
*Corresponding author.



*Proceedings of the 31st International Conference on Computational Linguistics*, pages 6798–6815
January 19–24, 2025. 

can be mainly divided into *training-free* (Leng et al., 2024; Huang et al., 2024a; Wang et al., 2024a; Manevich and Tsarfaty, 2024) and *training-based* approaches (Li et al., 2023a; Sun et al., 2023a; Zhao et al., 2023; Yu et al., 2024a; Gunjal et al., 2024; Liu et al., 2023; Yu et al., 2024b; Yue et al., 2024). Training-free approaches address potential hallucinations by postprocessing the outputs of MLLMs. While not requiring additional training costs, training-free approaches tend to reduce the inference speed. Training-based approaches primarily tackle hallucination issues by aligning MLLM outputs with human preferences. Direct Preference Optimization (DPO) (Rafailov et al., 2024) has emerged as a pivotal approach through annotating pairs of hallucinatory and non-hallucinatory responses and training alignment algorithms to improve the correspondence between images and text, where a *pair-response* consists of one non-hallucination and one hallucination.

The pair-response preference alignment method presents two critical challenges. First, the effectiveness of mitigating hallucinations largely depends on the quality of the pair-responses. Specifically, the description of an image may involve multiple dimensions. For example, when describing an image of a cat, it could include attributes like the cat's color, posture, and background. Additionally, the image may contain other objects, such as trees or toys. This complexity makes it difficult for the model to learn multi-dimensional information from the pair-responses. Second, obtaining such specific and diverse datasets of pair-responses poses a significant challenge. Current approaches primarily rely on human annotators or GPT-4V (Achiam et al., 2023) to label the outputs of MLLMs, which not only demands specialized knowledge but also incurs substantial time and financial costs.

To address the aforementioned challenges, we propose a RRHF-V framework, which automatically construct rank-responses to train the model for mitigating hallucinations in MLLMs. *First*, to obtain richer semantic information from responses, and inspired by RRHF (Yuan et al., 2024) (an algorithm for aligning LLMs that scores multiple sampled responses and learns to align the ranking of these responses with human preferences), we propose an alignment algorithm for MLLMs based on rank-responses. Instead of using pair-responses to train the model, we automatically construct $N$ hallucinatory responses for each non-hallucinatory response to form a *rank-response*, as illustrated in Figure 1. *Moreover*, unlike existing alignment algorithms like DPO for MLLMs and RRHF for LLMs that require multiple models to achieve preference alignment, our RRHF-V only needs to adjust the parameters of a single model. *Further*, to improve the model's ability to discern the differences between hallucinatory responses within a rank-response, we design a training objective based on rank loss and margin loss to balance these differences. *In addition*, we propose a method for automatically constructing rank-responses based on scene graphs, replacing the need for expensive and time-consuming human annotation of pair-responses. The main contributions of this paper are as follows:

- We are the first to propose using rank-responses to mitigate multimodal hallucinations. Compared to a pair-response for an image, the responses with different scores in a rank-response enable the model to learn rich semantic information across various dimensions of the image.
- We propose a scene graph-based approach to automatically construct rank-responses in a cost-effective and automatic manner.
- We design a novel training objective based on rank loss and margin loss to balance the differences between hallucinatory responses within a rank-response.

To validate the effectiveness of RRHF-V, we conducted experiments on four MLLM hallucination evaluation datasets, AMBER (Wang et al., 2023), POPE (Li et al., 2023b), MMhalBench (Sun et al., 2023b) and Object HalBench (Rohrbach et al., 2018) based on MLLMs of different sizes (LLaVA-1.5-7B (Liu et al., 2024c) and Tiny-LLaVA-1B (Zhou et al., 2024a)). Experimental results show that RRHF-V performs better than DPO in multimodal scenarios and effectively mitigates hallucinations. Further analysis reveals that rank-responses based on scene graphs also play a crucial role in mitigating hallucinations.

## 2 Related Work

### 2.1 Multimodal Large Language Models

Recent advancements in MLLMs research are primarily attributed to the evolution of LLMs (Wang et al., 2024b; Zhuo et al., 2024; Lu et al., 2024; Su et al., 2024). With the aid of advanced LLMs like LLaMA (Touvron et al., 2023) and Qwen (Bai

et al., 2023a), a batch of MLLMs such as LLaVA-1.5 (Liu et al., 2024c), Qwen-VL (Bai et al., 2023b), and mPLUGOwl2 (Ye et al., 2024) have emerged, which can comprehend and generate a wide array of content by utilizing information from distinct modalities like texts and images.

Despite the success, current MLLMs suffer from serious hallucination problems. Thus, in this paper, we focus on mitigating hallucination problems to promote the use of MLLMs in practical scenarios.

### 2.2 Hallucinations in MLLMs

Hallucinations in MLLMs have significantly impeded their usage in the real world, especially for tasks that rely on precise captions. Recently, numerous studies focus on the construction of datasets for *evaluating* hallucination phenomena (Rohrbach et al., 2018; Li et al., 2023b; Wang et al., 2023; Sun et al., 2023a; Zhong et al., 2024; Tong et al., 2024; Wu et al., 2024; Cao et al., 2024; Huang et al., 2024b; Mubarak et al., 2024; Jing et al., 2024; Lovenia et al., 2023; Zhai et al., 2023a; Wan and Bansal, 2022; Zhang et al., 2024b; Min et al., 2023; Yan et al., 2024). Concurrently, significant attention is directed towards *analyzing* the underlying causes of hallucinations (Tao et al., 2024; Sui et al., 2024; Fadeeva et al., 2024).

Moreover, various approaches have been proposed to *mitigate* hallucinations in MLLMs, including training-free and training-based approaches. *Training-free* approaches address potential hallucinations by post-processing the outputs of MLLMs (Leng et al., 2024; Huang et al., 2024a; Yin et al., 2023; Manevich and Tsarfaty, 2024; Wang et al., 2024a). For example, VCD (Leng et al., 2024) aims to address the model's over-reliance on linguistic priors and statistical biases by comparing the output distributions from unaltered and visually perturbed inputs. However, training-free approaches tend to reduce the inference speed. Instead, *training-based* approaches seek to mitigate hallucinations in MLLMs via further training, such as Supervised Fine-Tuning (SFT) (Liu et al., 2023) or preference learning (Sun et al., 2023a; Yu et al., 2024a; Li et al., 2023a; Zhao et al., 2023; Gunjal et al., 2024; Liu et al., 2024a; Yu et al., 2024b; Zhou et al., 2024b; Jiang et al., 2024; Jing and Du, 2024). For example, LRV (Liu et al., 2023) performs length controlled fine-tuning on visual instructions to mitigate hallucinations. LLaVA-RLHF (Sun et al., 2023a) is the first to train an MLLM to align with human preference. RLHF-V (Yu et al., 2024a) manually collects segment-level human preference and conduct dense direct preference optimization (DDPO) over the human feedback to reduce hallucinations. HA-DPO (Zhao et al., 2023) proposes a style-consistent DPO, which converts preference data pairs into a consistent format.

However, existing DPO-based methods typically use pair-responses (where a pair-response consists of one non-hallucinatory response and one hallucinatory response) to align MLLMs, addressing the hallucination problem. Moreover, most approaches heavily rely on manual annotation or GPT-4V (Achiam et al., 2023) when constructing pair-response datasets. Therefore, we aim to expand the number of hallucinatory responses and, for the first time, propose using rank-responses to mitigate multimodal hallucinations. Additionally, we introduce an automatic method for constructing rank-responses based on scene graphs, along with a training objective tailored for rank-responses.

## 3 Methodology

An overview of our proposed RRHF-V framework is shown in Figure 2, which consists of two main components: Rank-response Construction Pipeline and Training Objective.

### 3.1 Rank-response Construction Pipeline

#### 3.1.1 Data Source

We utilized the ShareGPT4V (Chen et al., 2023) dataset, which contains rich content and accurate textual descriptions. We randomly selected 5,000 *images* and their corresponding *questions* and textual *descriptions* (answers) from the ShareGPT4V dataset, denoted as $I$, $Q$, and $D$, respectively.

#### 3.1.2 Divide and Conquer

In the ShareGPT4V dataset, each image description typically consists of multiple sentences, a single pair-response construced from these sentences could hinder the model's ability to learn rich semantic information across various dimensions of the image. To address this, we introduce a fine-grained divide-and-conquer strategy that decomposes a description into atomic responses based on our designed *Divide-and-conquer Prompt* $T_D$ as detailed in Appendix E.1. Specifically, as illustrated in Figure 2 (Step ①), we employ the LLaMA3-8B model to exclude subjective statements and extract objective facts from the description $D_i$ ($D_i \in D$), and decompose them into a series of *atomic responses*

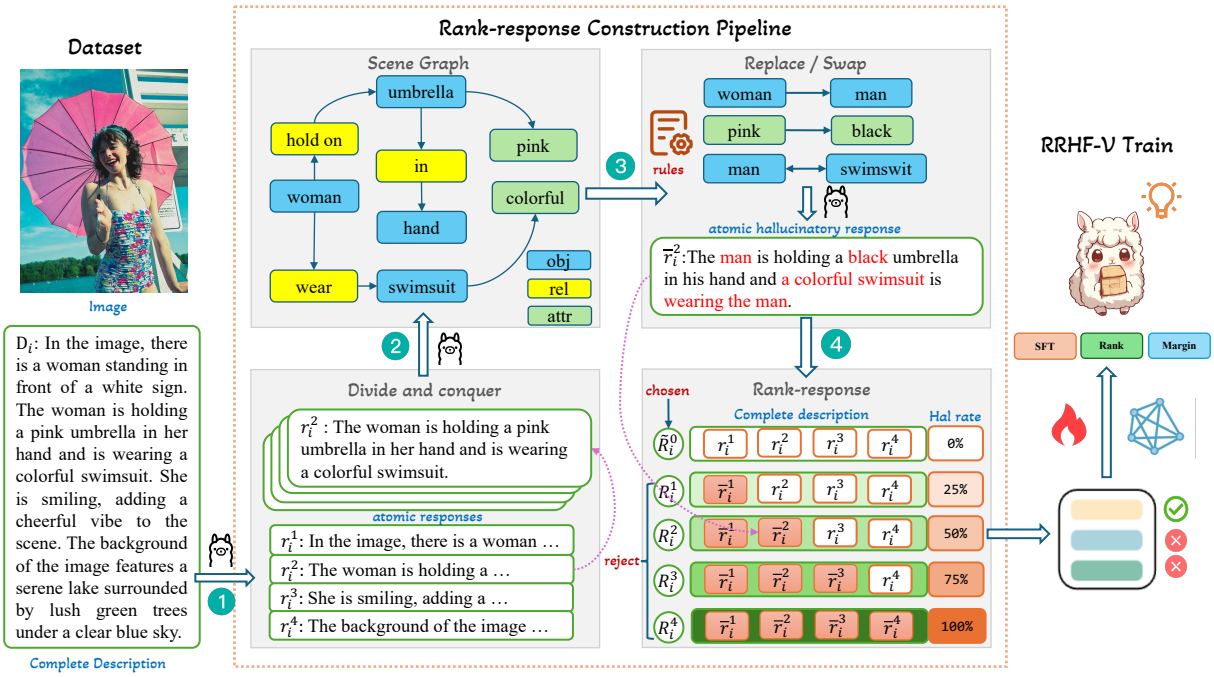



Figure 2: **Overview of the RRHF-V framework:** (1) Rank-response Construction Pipeline aims to automatically construct a rank-response for an image, which consists of four steps. Initially, the description of the image is decomposed into atomic responses (Step ①). Subsequently, for each atomic response, a corresponding scene graph is constructed (Step ②) and then further tuned to generate an atomic hallucinatory response (Step ③). Finally, by combining these atomic responses and atomic hallucinatory responses, a rank-response for the image, consisting of one non-hallucination $\tilde{R}_i^0$ and multiple ranking hallucinations $R_i^j$, is generated (Step ④). (2) We design a training objective based on SFT loss, rank loss, and margin loss to achieve the goal of mitigating hallucinations.

$r_i = \{r_i^1, r_i^2, \ldots, r_i^j\}$. These atomic responses convey complete semantic information independently, thereby enhancing the model's capability to understand fine-grained descriptions.

### 3.1.3 Scene Graph Generation

The hallucination issue in MLLMs primarily pertains to objects, their attributes, and the relationships among them (Zhou et al., 2024c). Considering that scene graphs are key components for understanding visual scenes, as they encapsulate rich semantic information, including objects, their attributes, and the relationships between them (Huang et al., 2024c). Therefore, we propose to generate a corresponding scene graph for each atomic response in order to construct high-quality hallucinatory responses.

Specifically, given an atomic response $r_i^j$ ($r_i^j \in r_i$), we parse it into a scene graph $G(r_i^j)$, where $G(r_i^j) = \langle O(r_i^j), R(r_i^j), E(r_i^j), A(r_i^j), K(r_i^j) \rangle$. Here, $O(r_i^j)$ is the set of objects mentioned in $r_i^j$, $R(r_i^j)$ is the set of relationship nodes, and $E(r_i^j) \subseteq O(r_i^j) \times R(r_i^j) \times O(r_i^j)$ is the set of hyper-edges representing actual relationships between objects. $K(r_i^j) \subseteq O(r_i^j) \times A(r_i^j)$ is the set of attribute pairs, where $A(r_i^j)$ is the set of attribute nodes associated with objects. To achieve this, in our framework we also use LLaMA3-8B model to parse atomic responses into scene graphs according to our carefully designed *Scene-graph-generation Prompt* $T_G$ as shown in Appendix E.2.

As shown in Figure 2 (Step ②), a scene graph for the atomic response $r_i^2$ is generated, where the objects, such as "woman", "umbrella" and "swimsuit" are the fundamental elements. The associated attributes, such as "pink" and "colorful" characterize the color or other attributes of objects. Relations such as "hold on" represent the spatial connections between objects.

### 3.1.4 Atomic Hallucinatory Response Generation

Our objective is to construct atomic hallucinatory responses with similar composition but different detailed semantics compared to atomic responses. Given an image-text pair ($I_i$ - $D_i$) ($I_i \in I, D_i \in D$) and a related scene graph $G(r_i^j)$ generated from the atomic response $r_i^j$ ($r_i^j \in r_i$), as shown in Figure

2 (Step ③), an *atomic hallucinatory response* $\overline{r}_i^j$ is generated via

$$\overline{r}_i^j = F(r_i^j, G(r_i^j)), \quad (1)$$

where $F$ is our defined sampling functions. Specifically, for triples $(object, relation, subject)$ in the scene graph, $\overline{r}_i^j$ is generated randomly through the following two sampling functions:

$$ObjSwap((O_1, Rel, O_2)) = (O_2, Rel, O_1), \quad (2)$$

$$ObjReplace((O_1, Rel, O_2)) = (O_3, Rel, O_2), \quad (3)$$

where $ObjSwap$ is the function to swap the object $O_1$ and the subject $O_2$ in the sentence, $Rel$ denote the relation. $ObjReplace$ is a function that replaces $O_1$ with $O_3$. $O_3$ is an entity that frequently co-occurs with $O_1$ in the real world.

For attribute pairs $(A_1, O_1)$ and $(A_2, O_2)$ in the scene graph, $\overline{r}_i^j$ is generated via

$$AttrSwap((A_1O_1), (A_2O_2)) = \begin{cases} (A_2O_1), (A_1O_2) & \text{if } O_1 \neq O_2, \\ pass & \text{if } O_1 = O_2, \end{cases} \quad (4)$$

where $AttrSwap$ is the function to exchange attributes between $(A_1, O_1)$ and $(A_2, O_2)$ in the scene graph.

#### 3.1.5 Rank-response Construction

As illustrated in Figure 2 (Step ④), we construct the final rank-responses. Formally, a *rank-response* $R_i$ constructed for an image-text pair ($I_i$ - $D_i$) consists of one non-hallucinatory response $\tilde{R}_i^0$ and $N$ hallucinatory responses $R_i^j$, where $N$ is the number of the atomic responses $r_i^j$ (where $j$ ranges from 1 to $N$) into which $D_i$ is divided. We further introduce the concept of "*hallucination rate*", where the hallucination rate $hr$ of each response represents the proportion of the atomic responses $r_i^j$ (Step ① in Figure 2) that are replaced by atomic hallucinatory responses $\overline{r}_i^j$ (Step ③ in Figure 2). Additionally, each response in the rank-response $R_i$ is assigned a score $S_i^k$, calculated as follows:

$$S_i^k = N \cdot (1 - hr_i^k), \quad (5)$$

where $S_i^k$, which represents the score corresponding to the $k$-th response in $R_i$ ($0 \leq k \leq N$), is inversely related to the hallucination rate. Ultimately, we obtain a rank-response for each image-text pair ($I_i$ - $D_i$) . The process of *rank-response construction* is summarized in Algorithm of Appendix A.

### 3.2 Training Objective

For an MLLM $\pi$, it starts from an input image $x_v$ and a question prompt $x_q$, drawn from $I$ and $Q$, respectively, and then generates a response $y$. Our target is to utilize an MLLM to produce an appropriate response $y$ with $x_v$ and $x_q$, while learning to distinguish hallucination differences by analyzing hallucination phenomena across multiple rank-responses with varying scores.

To achieve this, we introduce three loss functions: SFT loss, rank loss, and margin loss.

#### 3.2.1 SFT Loss

For a given $x_v$ and $x_q$, we expect model $\pi$ to generate a better response. Therefore, we require the model to learn the response with the highest score $S_i$. Consequently, the Supervised Fine-Tuning (SFT) loss is defined as follows:

$$i' = \arg\max_i S_i \quad (6)$$

$$L_{sft} = -\sum_t \log P_\pi(y_{i',t} | x_v, x_q, y_{i',<t}) \quad (7)$$

#### 3.2.2 Rank Loss

To align with scores $S$, we use our model $\pi$ to give an *implicit reward* $p_i$ for each $y_i$ by:

$$p_i = \frac{\sum_t \log P_\pi(y_{i,t} | x_v, x_q, y_{i,<t})}{\|y_i\|}, \quad (8)$$

where $p_i$ is conditional log probability (length-normalized) of $y_i$ under the model $\pi$. Following the idea of RRHF (Yuan et al., 2024), we also let the model $\pi$ give larger probabilities for better responses and give smaller probabilities for worse responses. We optimize this object by rank loss:

$$L_{rank} = \sum_{S_i < S_j} \max(0, p_i - p_j) \quad (9)$$

#### 3.2.3 Margin Loss

To further enhance the model's ability to discern the differences between hallucinatory responses within a rank-response, and to prevent the *implicit reward* difference between $y_i$ and $y_j$ within a rank-response from becoming excessively large or small, we introduce a margin loss

$$L_{margin} = \sum_{i=2}^{n} (p_0 - p_i - (i-1)\gamma)^2, \quad (10)$$

where $p_i$ denotes the *implicit reward* of the $i$-th response, and $\gamma$ is a constant offset.

| | Size | AMBER | | | | | | | | POPE | | MMHal Bench | Object HalBench | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | $Acc_e$ | $Acc_a$ | $Acc_s$ | $Acc_n$ | $Acc_{act}$ | $Acc_{all}$ | $F1_{all}$ | Score | $Acc_{adv}$ | $F1_{adv}$ | Score | $C_s \downarrow$ | $C_i \downarrow$ |
| VCD | 7B | 68.3 | 66.3 | 63.3 | 68.4 | 78.4 | 71.8 | 74.9 | 79.5 | 81.0 | 80.1 | 2.12 | 48.8 | 24.3 |
| OPERA | 7B | 77.8 | 26.3 | 29.6 | 22.8 | 15.9 | 75.2 | 78.3 | 36.2 | 79.1 | 86.4 | 2.15 | 45.1 | 22.3 |
| Less is more | 7B | 78.9 | 72.9 | 72.9 | 79.0 | 83.3 | 72.4 | 75.8 | 85.3 | 81.9 | 83.2 | 2.03 | 40.3 | 17.8 |
| LLaVA-RLHF | 7B | 64.5 | 72.7 | 71.1 | 72.1 | 83.7 | 68.7 | 74.4 | 83.1 | 75.8 | 78.9 | 1.80 | 46.2 | 24.5 |
| HA-DPO | 7B | 78.9 | 72.4 | 67.0 | 80.4 | 84.1 | 75.2 | 79.9 | 86.6 | 81.4 | 82.5 | 1.97 | 39.9 | 19.9 |
| FGAIF | 7B | - | - | - | - | - | - | - | - | 79.6 | 79.9 | 3.09 | 3.9 | 6.2 |
| LURE | 10B | - | - | - | - | - | 73.5 | 77.7 | - | - | - | 1.64 | 27.7 | 17.3 |
| Qwen-VL | 10B | 93.9 | 74.7 | 70.4 | 80.4 | 85.9 | 73.5 | 77.7 | 86.3 | 82.5 | 82.8 | 2.03 | 40.3 | 17.8 |
| RLHF-V | 13B | 92.1 | 78.3 | 77.0 | 77.9 | 87.0 | 72.6 | 75.0 | 84.3 | 81.8 | 79.4 | 2.81 | 12.2 | 7.5 |
| Tiny-LLaVA-1B | 1B | 68.9 | 66.3 | 62.2 | **72.4** | 74.9 | 65.0 | 63.9 | 79.6 | 78.8 | 80.6 | 1.56 | 61.4 | **32.5** |
| + DPO | 1B | 52.1 | 60.8 | 56.9 | 66.7 | 68.6 | 57.4 | 55.3 | 71.9 | 73.3 | 77.6 | **1.69** | 61.4 | 33.8 |
| **+ RRHF-V** | 1B | **83.6** | **69.5** | **66.0** | 72.1 | **83.5** | **70.5** | **77.6** | **83.1** | **82.6** | **82.1** | 1.38 | **58.5** | 33.7 |
| *APG* | - | 14.7 | 3.2 | 3.8 | -0.3 | 8.6 | 5.5 | 13.7 | 3.5 | 3.8 | 1.5 | -0.18 | -2.9 | 1.2 |
| *RPG* (%) | - | 21.34 | 4.83 | 6.11 | -0.41 | 11.48 | 8.46 | 21.44 | 4.40 | 4.82 | 1.86 | -11.54 | -4.72 | 3.69 |
| LLaVA-1.5-7B | 7B | 72.5 | 74.2 | 70.1 | 79.7 | 84.3 | 72.0 | 74.7 | 83.5 | 78.8 | 80.8 | **2.19** | 50.7 | 26.9 |
| + DPO | 7B | 53.5 | 69.3 | 66.0 | 72.2 | 82.1 | 64.2 | 64.1 | 78.6 | 72.8 | 77.6 | 2.05 | 52.1 | 26.9 |
| **+ RRHF-V** | 7B | **86.2** | **80.2** | **76.9** | **85.7** | **86.0** | **79.4** | **84.6** | **88.8** | **82.6** | **82.3** | 1.93 | **38.2** | **21.9** |
| *APG* | - | 13.7 | 6.0 | 6.8 | 6.0 | 1.7 | 7.4 | 9.9 | 5.3 | 3.8 | 1.5 | -0.26 | -12.5 | -5.0 |
| *RPG* (%) | - | 18.9 | 8.09 | 9.70 | 7.53 | 2.02 | 10.28 | 13.25 | 6.35 | 4.82 | 1.86 | -11.87 | -24.65 | -18.59 |

Table 1: **Main results** of Tiny-LLaVA-1B and LLaVA-1.5-7B trained with RRHF-V and base DPO. The best result for each metric in **each group** is in bold. APG and RPG indicate the absolute performance gains and the relative performance gains achieved by our model compared with the base model (Tiny-LLaVA-1B and LLaVA-1.5-7B). APG and RPG can be calculated by $APG = R_{rrhf-v} - R_{base}$ and $RPG = (R_{rrhf-v} - R_{base})/R_{base}$, where $R_{rrhf-v}$ and $R_{base}$ denote the results of our model and base model (Tiny-LLaVA-1B or LLaVA-1.5-7B), respectively. For reference, we also provide results for some typical methods using various MLLMs. The results on AMBER and POPE are our reproductions based on their original settings, while the results on MMHal and Object HalBench are from their original papers.

#### 3.2.4 RRHF-V Loss

The final loss is defined as the weighted sum of three individual loss functions:

$$L = L_{sft} + \alpha \cdot L_{rank} + \beta \cdot L_{margin}, \qquad (11)$$

where $\alpha$ and $\beta$ denote the hyperparameters.

## 4 Experiment

### 4.1 Experimental Setup

#### 4.1.1 Models

We apply RRHF-V on two multimodal large models in different sizes, Tiny-LLaVA-1B (Zhou et al., 2024a) and LLaVA-1.5-7B (Liu et al., 2024b). More details about these MLLMs can be found in Appendix B.1.

#### 4.1.2 Training Data

We sample 5K data from ShareGPT4V (Chen et al., 2023) for constructing rank-responses and training. More details about ShareGPT4V can be found in Appendix B.2.

#### 4.1.3 Evaluation Benchmarks

We evaluate the performance of RRHF-V on four widely used benchmarks (POPE (Li et al., 2023b), AMBER (Wang et al., 2023), MMHalBench (Sun et al., 2023b), Object HalBench (Rohrbach et al., 2018)) for MLLMs with a special focus on hallucination. The benchmarks are detailed in Appendix B.3.

#### 4.1.4 Baselines

We primarily compare RRHF-V with standard DPO. We also provide the results of general leading-edge MLLMs (Tiny-LLaVA-1B (Zhou et al., 2024a), LLaVA-v1.5-7B (Liu et al., 2024b) and Qwen-VL-Chat-10B (Bai et al., 2023b)) and several *training-free* approaches (VCD (Leng et al., 2024), OPERA (Huang et al., 2024a)). Regarding the *training-based* approaches, we select some typical methods, including FGAIF (Jing and Du, 2024), Less is more (Yue et al., 2024), LLaVA-RLHF (Sun et al., 2023a) and HA-DPO (Zhao et al., 2023) under the same model size, as well as methods based on larger model sizes, such as RLHF-V (Yu et al.,

2024a) and LURE (Zhou et al., 2024c).

#### 4.1.5 Implementation Details

We conducted experiments on both the Tiny-LLaVA-1B and LLaVA-1.5-7B models. More details can be found in Appendix B.4.

### 4.2 Main Results

Table 1 presents the primary experimental results. We observe the following points:

- Across four benchmark tests, both LLaVA-1.5-7b and Tiny-LLaVA-1b generally outperform the DPO method, indicating that the rank-response strategy by extending the number of hallucinatory responses significantly enhances model performance.
- LLaVA-1.5-7B achieved improvements of 13.25% and 4.82% on AMBER($F1_{all}$) and POPE($ACC_{adv}$). Similarly, Tiny-LLaVA-1B achieved increases of 21.44% and 21.34% on AMBER($F1_{all}$) and AMBER($F1_e$). These results demonstrate the effectiveness of RRHF-V in mitigating hallucinations of MLLMs.
- Furthermore, RRHF-V consistently outperforms a range of training-free and training-based methods at the same model size (7B), and also demonstrates superior performance compared to the larger model (10B) across most metrics. These results further demonstrate the effectiveness of our approach.

### 4.3 Analysis

We conduct analysis on RRHF-V considering the following questions: (**Q1**) Has the introduction of scene graphs effectively enhanced the quality of rank-responses? (**Q2**) Are all three loss functions in the training objective necessary? (**Q3**) How does RRHF-V's performance scale with feedback data amount? (**Q4**) Does the number of rank-responses have a significant impact on the final results? (**Q5**) How does the overhead of the rank-response data construction pipeline proposed in RRHF-V? (**Q6**) How do the training time and memory overhead of the RRHF-V compare to DPO?

**A1: Rank-responses constructed based on scene graphs is better.** To validate the effectiveness of rank-responses constructed from scene graphs, we employ the Llama3-8B to directly generate atomic hallucinatory responses from atomic responses, rather than through scene graph transformations. The results are presented in Table 2.

| | POPE | | AMBER | |
|---|---|---|---|---|
| | ACC | F1 | ACC | F1 |
| **w** Scene Graph | **82.6** | **82.3** | **79.4** | **84.6** |
| **w/o** Scene Graph | 77.7 | 80.3 | 73.8 | 77.0 |

Table 2: **Comparison of generated rank-responses depending on whether scene graphs are used.** The results indicate that the RRHF-V's performance significantly deteriorates in the absence of scene graphs.

| | POPE | | AMBER | |
|---|---|---|---|---|
| | ACC | F1 | ACC | F1 |
| $L_{sft}$ | 80.4 | 80.4 | 74.8 | 79.1 |
| $L_{sft} + L_{rank}$ | 80.6 | 81.2 | 76.3 | 81.6 |
| $L_{sft} + L_{margin}$ | 79.7 | 81.3 | 75.9 | 79.6 |
| $L_{rank} + L_{margin}$ | 79.1 | 81.0 | 74.7 | 77.7 |
| $L_{sft} + L_{rank} + L_{margin}$ | **82.6** | **82.3** | **79.4** | **84.6** |

Table 3: **Ablation results of different loss functions in RRHF-V.** The results indicate that the sft loss, rank loss, and margin loss each play an indispensable role in achieving the training objectives.

We observe that the results are not ideal without using scene graphs. We suggest that directly converting atomic responses into atomic hallucinatory responses introduces uncertainty, hindering effective learning. This further validates the effectiveness of our proposed method for constructing rank-responses based on scene graphs.

**A2: Each component in our loss enhances model performance.** To validate the contribution of each loss function to model performance, we conduct an ablation study in Table 3. The results demonstrate that all three loss functions are essential for achieving the training objectives. They complement each other, and their combined use is crucial for improving overall model performance. More detailed experiments on hyperparameters $\alpha$, $\beta$, and $\gamma$ in the loss can be found in Appendix C.

**A3: Scaling feedback data leads to promising results.** We sample 5k data from ShareGPT4V as mentioned in Section 4.1.2 for training in our main implementation. To further ensure fairness in data selection, we randomly sample 10,000 instances from the COCO image dataset within ShareGPT4V and construct datasets of varying sizes: 10k, 5k, 2k, and 1k. Rank-responses are constructed based on these datasets. As shown in Table 4, hallucinations are effectively mitigated as the data scale increases. Based on this trend, we conjecture that further expanding data scale may lead to additional improvements in model performance.

| | POPE | | AMBER | |
|---|---|---|---|---|
| | ACC | F1 | ACC | F1 |
| coco-1k | 77.2 | 79.9 | 71.9 | 74.3 |
| coco-2k | 77.3 | 80.0 | 73.5 | 76.2 |
| coco-5k | 80.9 | 81.7 | 75.7 | 80.8 |
| coco-10k | **83.3** | **82.2** | **76.8** | **83.9** |

Table 4: **Comparison of different amounts of feedback data.** A set of 10k image-text pairs was selected from the COCO dataset and divided into four different data scales for training. The results indicate that model performance improves as the data scale increases.

| | POPE | | AMBER | |
|---|---|---|---|---|
| | ACC | F1 | ACC | F1 |
| Number = 1 | 78.2 | 80.5 | 71.2 | 73.2 |
| Number = 2 | 78.9 | 81.1 | 74.2 | 77.0 |
| Number = 3 | 80.4 | 81.5 | 76.0 | 80.0 |
| Number = 4 | 81.2 | 81.7 | 77.1 | 82.3 |
| Number = 5 | **82.6** | **82.3** | **79.4** | **84.6** |
| Number = 10 | 78.9 | 80.9 | 55.6 | 60.3 |

Table 5: **Comparison of the number of hallucinatory responses in a rank-response.** The results show that as the number increases, the RRHF-V's performance gradually improves; however, when the number becomes too large, performance begins to deteriorate.

**A4: Rank-responses necessitate a reasonable number.** Although we observe that the occurrence of hallucination phenomena decreases with the increase in the number of rank-responses, a higher number of hallucinatory responses are not necessarily beneficial for model performance. As shown in Table 5, an excessive number of hallucinatory responses may actually lead to a decline in model performance during training.

**A5: The cost of rank-response construction may be acceptable.** As illustrated in Figure 3, we present the time cost for generating 5000 rank-responses using the LLaMA3-8B model. Compared to manually annotated response methods, our approach is automated. Moreover, due to limitations in resources and data scale, we do not provide a direct comparison of the time cost for generating a similar scale of rank-responses using the GPT-4V. Nevertheless, the time cost of generating rank-responses with open-source models remains within an acceptable range for practical applications.

**A6: RRHF-V lags behind DPO in LoRA, but excels in full parameter fine-tuning.** Figure 4(a) illustrates the results of applying the LoRA method to train the LLaVA-1.5-7B model. The experimental results indicate that RRHF-V has slightly

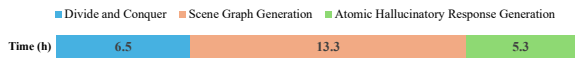



Figure 3: The time overhead for each step involved in constructing 5000 rank-responses using the LLaMA3-8B model.

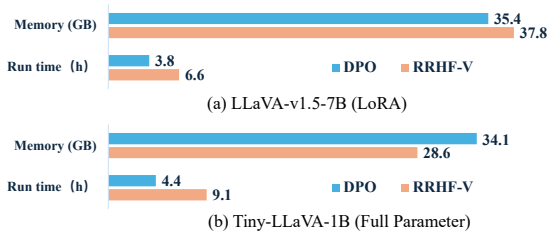



Figure 4: Figure (a) and Figure (b) respectively illustrate the performance and computational overhead of the LLaVA-1.5-7B model with LoRA fine-tuning using DPO and RRHF-V, and the Tiny-LLaVA-1B model with full parameter fine-tuning, under identical training parameters and dataset scales.

higher memory usage and training time compared to DPO, suggesting that even with the extended rank-responses, the computational overhead increase is minimal. Howerve, our RRHF-V only needs to adjust the parameters of a single model, unlike DPO which need multiple models. This advantage is particularly prominent in full parameter fine-tuning. Due to hardware resource constraints, we performed full parameter fine-tuning on the Tiny-LLaVA-1B model. Figure 4(b) illustrates that RRHF-V outperforms DPO in terms of memory usage. Based on these results, we anticipate that for larger models with full parameter fine-tuning, RRHF-V will further excel in memory consumption, thus enhancing its practical advantages. We will conduct more extensive exploration and experimentation in future work.

### 4.4 Case Study

To provide a more intuitive demonstration of RRHF-V's performance in mitigating hallucinations, we conducted case studies of the RRHF-V and the original LLaVA-1.5-7B model. For detailed results, please refer to the Appendix D.

## 5 Conclusion

In this paper, we present RRHF-V, a novel approach to mitigate multimodal hallucinations in MLLMs. By using rank-responses instead of pair-responses, we expand the number of hallucinatory responses,

enabling the model to capture richer semantic information of an image-text input. We further propose a scene graph-based approach to automatically construct high-quality rank-responses, avoiding manual annotation. We also design a training objective tailed for training the models on rank-responses. Experiments on two MLLMs and four benchmarks show that RRHF-V effectively mitigates hallucinations, outperforming traditional pair-response methods like DPO.

## Limitations

Firstly, although extensive experiments have demonstrated the superior effectiveness of RRHF-V over DPO, we have yet to combine it with other DPO-based improvement methods. The analysis of such combined approaches will be left for future work. Secondly, our exploration of the number of hallucinatory responses remains insufficient. We believe that the optimal approach should adaptively determine the appropriate number of hallucinatory responses based on the data rather than relying on pre-defined values. In future work, we will further investigate methods for optimally determining the number of hallucinatory responses. Lastly, our current data construction pipeline is limited to image description data and has not yet been extended to tasks such as visual reasoning or visual question answering. In the future, we will explore how to extend RRHF-V to more complex and diverse data types.

**Acknowledgments**. The authors sincerely thank the reviewers for their valuable comments, which improved the paper. The work is supported by the National Natural Science Foundation of China (62276057), and Sponsored by CAAI-MindSpore Open Fund, developed on OpenI Community.

## References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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

Kevin Black, Michael Janner, Yilun Du, Ilya Kostrikov, and Sergey Levine. 2024. Training diffusion models with reinforcement learning. In *The Twelfth International Conference on Learning Representations*.

Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402.

Qingxing Cao, Junhao Cheng, Xiaodan Liang, and Liang Lin. 2024. VisDiaHalBench: A visual dialogue benchmark for diagnosing hallucination in large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Lin Chen, Jisong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multimodal models with better captions. *arXiv preprint arXiv:2311.12793*.

Wei-Lin Chiang, Zhuohan Li, Zi Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, et al. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna.lmsys.org (accessed 14 April 2023)*, 2(3):6.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Hamdy Mubarak, Evgenii Tsymbalov, Gleb Kuzmin, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, and Maxim Panov. 2024. Fact-checking the output of large language models via token-level uncertainty quantification. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9367–9385, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Zigang Geng, Binxin Yang, Tiankai Hang, Chen Li, Shuyang Gu, Ting Zhang, Jianmin Bao, Zheng Zhang, Houqiang Li, Han Hu, et al. 2024. Instructdiffusion: A generalist modeling interface for vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12709–12720.

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

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. 2024a. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13418–13427.

Wen Huang, Hongbin Liu, Minxin Guo, and Neil Gong. 2024b. Visual hallucinations of multi-modal large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 9614–9631, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yufeng Huang, Jiji Tang, Zhuo Chen, Rongsheng Zhang, Xinfeng Zhang, Weijie Chen, Zeng Zhao, Zhou Zhao, Tangjie Lv, Zhipeng Hu, et al. 2024c. Structure-clip: Towards scene graph knowledge to enhance multi-modal structured representations. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 2417–2425.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. 2024. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 27036–27046.

Liqiang Jing and Xinya Du. 2024. Fgaif: Aligning large vision-language models with fine-grained ai feedback. *arXiv preprint arXiv:2404.05046*.

Liqiang Jing, Ruosen Li, Yunmo Chen, and Xinya Du. 2024. Faithscore: Fine-grained evaluations of hallucinations in large vision-language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 5042–5063.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. 2023. Segment anything. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 4015–4026.

Xin Lai, Zhuotao Tian, Yukang Chen, Yanwei Li, Yuhui Yuan, Shu Liu, and Jiaya Jia. 2024. Lisa: Reasoning segmentation via large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9579–9589.

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

Chunyuan Li, Cliff Wong, Sheng Zhang, Naoto Usuyama, Haotian Liu, Jianwei Yang, Tristan Naumann, Hoifung Poon, and Jianfeng Gao. 2024. Llava-med: Training a large language-and-vision assistant for biomedicine in one day. *Advances in Neural Information Processing Systems*, 36.

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

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. 2023b. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. 2023. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *The Twelfth International Conference on Learning Representations*.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024b. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2024c. Visual instruction tuning. *Advances in neural information processing systems*, 36.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. 2023. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. *arXiv preprint arXiv:2310.05338*.

Zhenyi Lu, Jie Tian, Wei Wei, Xiaoye Qu, Yu Cheng, Wenfeng Xie, and Dangyang Chen. 2024. Mitigating boundary ambiguity and inherent bias for text classification in the era of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*.

Avshalom Manevich and Reut Tsarfaty. 2024. Mitigating hallucinations in large vision-language models (LVLMs) via language-contrastive decoding (LCD). In *Findings of the Association for Computational Linguistics ACL 2024*, pages 6008–6022, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yu Meng, Mengzhou Xia, and Danqi Chen. 2024. Simpo: Simple preference optimization with a reference-free reward. *arXiv preprint arXiv:2405.14734*.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Wei Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2023. Factscore: Fine-grained atomic evaluation of factual precision in long form text generation. *arXiv preprint arXiv:2305.14251*.

Hamdy Mubarak, Hend Al-Khalifa, and Khaloud Suliman Alkhalefah. 2024. Halwasa: Quantify and analyze hallucinations in large language models: Arabic as a case study. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8008–8015, Torino, Italia. ELRA and ICCL.

OpenAI. 2023. Chatgpt: optimizing language models for dialogue. openai. 2022. *URL https://openai.com/blog/chatgpt*.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Alizée Pace, Jonathan Mallinson, Eric Malmi, Sebastian Krause, and Aliaksei Severyn. 2024. West-of-n: Synthetic preference generation for improved reward modeling. *arXiv preprint arXiv:2401.12086*.

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

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. 2024. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. 2018. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, Brussels, Belgium. Association for Computational Linguistics.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Oleksii Sidorov, Ronghang Hu, Marcus Rohrbach, and Amanpreet Singh. 2020. Textcaps: a dataset for image captioning with reading comprehension. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part II 16*, pages 742–758. Springer.

Zhaochen Su, Juntao Li, Jun Zhang, Tong Zhu, Xiaoye Qu, Pan Zhou, Yan Bowen, Yu Cheng, et al. 2024. Living in the moment: Can large language models grasp co-temporal reasoning? *arXiv preprint arXiv:2406.09072*.

Peiqi Sui, Eamon Duede, Sophie Wu, and Richard So. 2024. Confabulation: The surprising value of large language model hallucinations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14274–14284, Bangkok, Thailand. Association for Computational Linguistics.

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

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

Mingxu Tao, Quzhe Huang, Kun Xu, Liwei Chen, Yansong Feng, and Dongyan Zhao. 2024. Probing multimodal large language models for global and local semantic representations. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 13050–13056, Torino, Italia. ELRA and ICCL.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Shengbang Tong, Zhuang Liu, Yuexiang Zhai, Yi Ma, Yann LeCun, and Saining Xie. 2024. Eyes wide shut? exploring the visual shortcomings of multimodal llms. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9568–9578.

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

David Wan and Mohit Bansal. 2022. Evaluating and improving factuality in multimodal abstractive summarization. *arXiv preprint arXiv:2211.02580*.

Junyang Wang, Yuhang Wang, Guohai Xu, Jing Zhang, Yukai Gu, Haitao Jia, Ming Yan, Ji Zhang, and Jitao Sang. 2023. An llm-free multi-dimensional benchmark for mllms hallucination evaluation. *arXiv preprint arXiv:2311.07397*.

Xintong Wang, Jingheng Pan, Liang Ding, and Chris Biemann. 2024a. Mitigating hallucinations in large vision-language models with instruction contrastive decoding. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 15840–15853, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Yue Wang, Tianfan Fu, Yinlong Xu, Zihan Ma, Hongxia Xu, Bang Du, Yingzhou Lu, Honghao Gao, Jian Wu, and Jintai Chen. 2024b. Twin-gpt: Digital twins for clinical trials via large language model. *ACM Transactions on Multimedia Computing, Communications and Applications*.

Mingrui Wu, Jiayi Ji, Oucheng Huang, Jiale Li, Yuhang Wu, Xiaoshuai Sun, and Rongrong Ji. 2024. Evaluating and analyzing relationship hallucinations in large vision-language models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 53553–53570. PMLR.

Bowen Yan, Zhengsong Zhang, Liqiang Jing, Eftekhar Hossain, and Xinya Du. 2024. Fiha: Autonomous hallucination evaluation in vision-language models with davidson scene graphs. *arXiv preprint arXiv:2409.13612*.

Kevin Yang, Dan Klein, Asli Celikyilmaz, Nanyun Peng, and Yuandong Tian. 2023. Rlcd: Reinforcement learning from contrast distillation for language model alignment. *arXiv preprint arXiv:2307.12950*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multimodal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051.

Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. 2023. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*.

Tianyu Yu, Yuan Yao, Haoye Zhang, Taiwen He, Yifeng Han, Ganqu Cui, Jinyi Hu, Zhiyuan Liu, Hai-Tao Zheng, Maosong Sun, et al. 2024a. Rlhf-v: Towards trustworthy mllms via behavior alignment from fine-grained correctional human feedback. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13807–13816.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, et al. 2024b. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Hongyi Yuan, Zheng Yuan, Chuanqi Tan, Wei Wang, Songfang Huang, and Fei Huang. 2024. Rrhf: Rank responses to align language models with human feedback. *Advances in Neural Information Processing Systems*, 36.

Zihao Yue, Liang Zhang, and Qin Jin. 2024. Less is more: Mitigating multimodal hallucination from an EOS decision perspective. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Bohan Zhai, Shijia Yang, Xiangchen Zhao, Chenfeng Xu, Sheng Shen, Dongdi Zhao, Kurt Keutzer, Manling Li, Tan Yan, and Xiangjun Fan. 2023a. Halleswitch: Rethinking and controlling object existence hallucinations in large vision language models for detailed caption. *arXiv preprint arXiv:2310.01779*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023b. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024a. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Shilong Zhang, Peize Sun, Shoufa Chen, Min Xiao, Wenqi Shao, Wenwei Zhang, Yu Liu, Kai Chen, and Ping Luo. 2023. Gpt4roi: Instruction tuning large language model on region-of-interest. *arXiv preprint arXiv:2307.03601*.

Yue Zhang, Jingxuan Zuo, and Liqiang Jing. 2024b. Fine-grained and explainable factuality evaluation for multimodal summarization. *arXiv preprint arXiv:2402.11414*.

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

Weihong Zhong, Xiaocheng Feng, Liang Zhao, Qiming Li, Lei Huang, Yuxuan Gu, Weitao Ma, Yuan Xu, and Bing Qin. 2024. Investigating and mitigating the multimodal hallucination snowballing in large vision-language models. In *Proceedings of the 62nd*

*Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Baichuan Zhou, Ying Hu, Xi Weng, Junlong Jia, Jie Luo, Xien Liu, Ji Wu, and Lei Huang. 2024a. Tinyllava: A framework of small-scale large multimodal models. *arXiv preprint arXiv:2402.14289*.

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

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

Linhai Zhuo, Yuqian Fu, Jingjing Chen, Yixin Cao, and Yu-Gang Jiang. 2024. Unified view empirical study for large pretrained model on cross-domain few-shot learning. *ACM Transactions on Multimedia Computing, Communications and Applications*.

## A Algorithm of Constructing Rank-responses

The pseudocode for Rank-response Construction is shown in Algorithm 1.

**Algorithm 1:** Rank-Response Construction

**Input:**
$\mathcal{D}$: Set of image-text pairs $\{(I_i, D_i)\}$ where $i = 1, 2, \ldots, K$
$N$: Number of atomic responses $r_i^j$ into which each $D_i$ is divided
**Output:** $\mathcal{R}$: Set of rank-responses $\{R_i\}$ with corresponding scores $\{S_i^k\}$

$\mathcal{R} \leftarrow \emptyset$
**for** *each image-text pair* $(I_i, D_i) \in \mathcal{D}$ **do**
  Divide $D_i$ into $N$ atomic responses: $D_i = \{r_i^1, r_i^2, \ldots, r_i^N\}$;
  **for** *each atomic response* $r_i^j \in D_i$ **do**
    Construct corresponding atomic hallucinatory response $\overline{r}_i^j$;
  Initialize $R_i = \{\tilde{R}_i^0, R_i^1, R_i^2, \ldots, R_i^N\}$, where $\tilde{R}_i^0$ is the non-hallucinatory response, and $R_i^j$ are hallucinatory responses;
  **for** $k = 0$ *to* $N$ **do**
    $R_i^k$ represents a description with some atomic responses $r_i^j$ replaced by their corresponding hallucinatory responses $\overline{r}_i^j$;
    Calculate the hallucination rate $\text{hr}_i^k$ as the proportion of atomic responses in $R_i^k$ that are replaced by hallucinatory responses;
    Compute score $S_i^k$ for $R_i^k$:

$$S_i^k = N \cdot (1 - \text{hr}_i^k)$$

  $\mathcal{R} \leftarrow \mathcal{R} \cup \{(R_i, \{S_i^k\})\}$;
return $\mathcal{R}$

## B Evaluation Details

We introduce more evaluation details, including baseline models and evaluation benchmarks.

### B.1 Models

- Tiny-LLaVA-1B (Zhou et al., 2024a) is a 1.1B model building upon SigLIP (Zhai et al., 2023b) and TinyLlama (Zhang et al., 2024a). It is pretrained on 1.8M image-text pairs and finetuned on 1.3M instruction tuning data.
- LLaVA-1.5-7B (Liu et al., 2024b) is a 7B model based on CLIP (Radford et al., 2021) and Vincuna (Chiang et al., 2023). It is pretrained on 558K image-text pairs and finetuned on 665K instruction tuning data.

### B.2 Datasets

The ShareGPT4V(Chen et al., 2023) dataset is a comprehensive and diverse collection of image descriptions specifically gathered for supervised fine-tuning, utilizing the GPT-4 Vision (Achiam et al., 2023). This dataset comprises approximately 100,000 images compiled from various data sources, including images for detection (Lin et al., 2014) and segmentation (Kirillov et al., 2023), images containing complex text (Sidorov et al., 2020), and various web images featuring artworks, landmarks, celebrities, etc. (Schuhmann et al., 2021; Sharma et al., 2018; Ordonez et al., 2011).

In the original ShareGPT4V text, it is explained that for each image, carefully designed data-specific prompts to guide GPT-4 Vision in producing detailed descriptions that account for world knowledge, object attributes, spatial relationships, and aesthetic assessments. To maintain the quality and consistency of these descriptions, a general prompt structure is established for basic descriptions, with additional specific prompts tailored for each data source. These specific prompts are designed to incorporate relevant knowledge about the images, such as the names and geographical locations of landmarks, ensuring the descriptions go beyond superficial appearances. For example, the Eiffel Tower should not merely be described as a "tall iron tower", and a photo of Einstein should not simply be summarized as "an old man".

Furthermore, some images included prompts related to aesthetics to enhance the comprehensiveness of the descriptions. The careful design of these prompts and the high-quality image descriptions generated by GPT-4 Vision make the ShareGPT4V dataset a valuable resource for fine-tuning visual language models.

### B.3 Evaluation Benchmarks

We introduce additional details about the benchmarks we used for evaluation.

- POPE (Li et al., 2023b) is a mainstream dataset for hallucination evaluation

in MLLMs, contains 9,000 questions of 3 types. POPE targets at object existence of fixed categories (80 COCO) in images, supplying Yes/No response. The model's accuracy is benchmarked against the ground truth answer. We use adversarial questions and report the accuracy and F1 metric.

- AMBER (Wang et al., 2023) is a multi-dimensional hallucination benchmark comprising more than 15k samples. We focus on the discriminative task and report the accuracy, F1 metric and AMBER Score.

- MMHalBench (Sun et al., 2023b) is a practical question answering benchmark containing eight question categories and 12 object topics. Following the official setting, we use GPT-4 (Achiam et al., 2023) to assess the overall quality of response with a score between zero and six.

- Object HalBench (Rohrbach et al., 2018) is a widely adopted benchmark to assess object hallucination. We report the CHAIR scores (Rohrbach et al., 2018) assessing hallucination rate of response level ($CHAIR_s$) and object level ($CHAIR_i$), with GPT-3.5 participating in the evaluation.

### B.4 Implementation Details

For the construction of rank-responses, we employed the LLaMA3-8B-PF16 model. The entire data generation process produced a total of 5k data, with a completion time of less than 26 hours. The detailed breakdown of each step's cost is illustrated in Figure 3.

For the training of RRHF-V, we set N (the number of hallucinatory responses) to 5 and use 5k data in total. We primarily compare RRHF-V with standard DPO. The DPO baseline shares the same training process, data (we consider that DPO performs best when training on responses with large gaps (Yang et al., 2023; Meng et al., 2024; Pace et al., 2024), so we selected the highest-scoring and lowest-scoring responses to form a pair-response), and hyper-parameters, despite having different learning objectives.

- RRHF-V for LLaVA-1.5-7B. We train it for 2 epochs with LoRA (Hu et al., 2022) where lora rank is 100 and lora alpha is 50. Learning rate is 1.4e-5 and batch size is 1 and the hyperparameter $\alpha$ is 100, $\beta$ is 100, and $\gamma$ is 0.4.

- DPO for LLaVA-1.5-7B. We train it for 2 epochs with LoRA where lora rank is 100 and lora alpha is 50. Learning rate is 1.4e-5 and batch size is 1.

- RRHF-V for Tiny-LLaVA-1B. Due to the instability of small-sized models, we adopted a two-stage training strategy for Tiny-LLaVA-1B. In the first stage, We train it for 5 epochs with LoRA where lora rank is 256 and lora alpha is 128. Learning rate is 1.4e-5 and batch size is 3 and the hyperparameter $\alpha$ is 100, $\beta$ is 100, and $\gamma$ is 0.4. In the second stage, we only change $\gamma$ to 0.1 and train it for 2 epochs.

- DPO for Tiny-LLaVA-1B. We train it for 2 epochs with LoRA where lora rank is 256 and lora alpha is 128. Learning rate is 1.4e-5 and batch size is 3.

| Hyperparameters | | | POPE | | AMBER | |
|---|---|---|---|---|---|---|
| $\alpha$ | $\beta$ | $\gamma$ | ACC | F1 | ACC | F1 |
| 50 | 20 | 0.4 | **82.8** | 82.1 | 74.1 | 84.0 |
| 50 | 50 | 0.4 | 82.3 | 82.3 | 78.0 | 83.1 |
| 50 | 50 | 0.5 | 82.3 | **82.5** | 77.4 | 82.9 |
| 50 | 50 | 0.6 | 81.8 | 82.2 | 78.9 | 83.6 |
| 100 | 100 | 0.35 | 80.8 | 81.8 | 76.5 | 80.8 |
| 100 | 100 | 0.4 | 82.7 | 82.3 | **79.4** | **84.6** |
| 100 | 100 | 0.5 | 81.3 | 81.9 | 77.7 | 82.2 |
| 150 | 150 | 0.4 | 81.2 | 82.0 | 76.8 | 81.5 |
| 200 | 200 | 0.4 | 79.9 | 81.4 | 76.9 | 80.9 |

Table 6: **Comparison of different hyperparameters $\alpha$, $\beta$, $\gamma$.** The results show that different values of hyperparameters $\alpha$, $\beta$, $\gamma$ have a significant impact on the experimental results.

## C Additional Experiments of hyperparameters $\alpha$, $\beta$, and $\gamma$ in the Loss

To systematically evaluate the impact of hyperparameters $\alpha$, $\beta$, and $\gamma$ on model performance, we conducted ablation experiments as shown in Table 6. Through a series of experimental analyses, we systematically compared the effects of different values of $\alpha$, $\beta$, and $\gamma$ on the loss function. Based on the experimental results, we ultimately determined $\alpha = 100$, $\beta = 100$, and $\gamma = 0.4$ as the final hyperparameter configuration for the loss function.

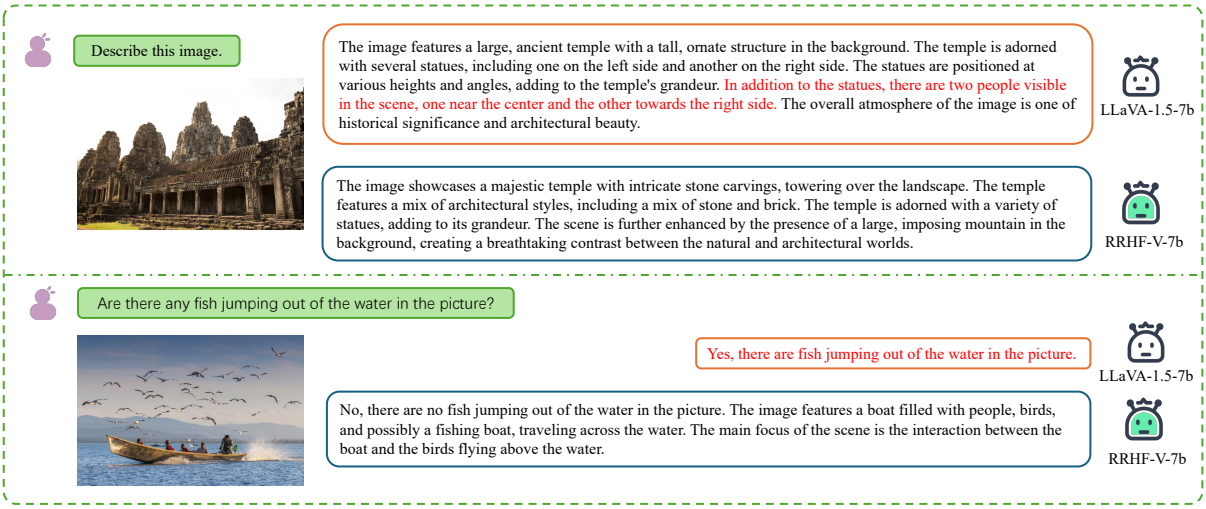



Figure 5: **RRHF-V's performance in mitigating hallucinations.** We compared the performance of RRHF-V with the original LLaVA-1.5-7B model. Specifically, the top section of the image illustrates the model's ability to provide detailed descriptions of visual information, while the bottom section evaluates the model's accuracy in recognizing co-occurring entities—an area particularly prone to hallucinations. Text highlighted in **red** denotes hallucinatory descriptions.

# D Case Study

Details of RRHF-V's performance in mitigating hallucinations are shown in Figure 5. We present two case studies involving the RRHF-V model: (1) In the upper part, the original LLaVA-1.5-7B model exhibits hallucinations when describing the image, incorrectly identifying two people in front of a temple, whereas the RRHF-V accurately depicts the image content without hallucinations. (2) In the lower part, we compare the ability of RRHF-V and the original LLaVA-1.5-7B to recognize entities within the image. We introduced the hallucination of "fish", which is most likely to co-occur with entities in the image. The LLaVA-1.5-7B mistakenly identified the presence of "fish" in the image, while RRHF-V remained unaffected by the co-occurrence information and made the correct judgment. These findings further support the earlier conclusion that RRHF-V effectively mitigates hallucination in MLLMs.

# E Prompts in Rank-response Construction Pipeline

## E.1 Details of Divide-and-conquer Prompt Templates

Details of our Divide-and-conquer prompt template are shown below in Figure 6.

## E.2 Details of Scene-graph-generation Prompt Templates

Details of our Scene-graph-generation prompt template are shown below in Figure 7.

**Task:**
Given a paragraph, divide it into 5 sentences.

**For example:**
***input:*** *{*
*"In the tranquil setting of a park, a row of six neatly trimmed spherical bushes stand in line, their vibrant green color contrasting beautifully with the verdant grass beneath them. On the left foreground, a red fire hydrant adds a pop of color to the scene. In the distance, a building and a tree can be seen against the backdrop of a clear blue sky. The image captures the essence of a peaceful day in the park."*
*}*
***output:****{*
*"1": "In the tranquil setting of a park, a row of six neatly trimmed spherical bushes stand in line.",*
*"2": "Their vibrant green color contrasts beautifully with the verdant grass beneath them.",*
*"3": "On the left foreground, a red fire hydrant adds a pop of color to the scene.",*
*"4": "In the distance, a building and a tree can be seen against the backdrop of a clear blue sky.",*
*"5": "The image captures the essence of a peaceful day in the park."*
*}*

**Require:**
1. Divide the paragraph into 5 sentences based on the overall semantics.
2. Excluding subjective statements.
3. If the number of sentences exceeds 5, merge some sentences to make it 5 sentences.
4. If the number of sentences is less than 5, split some sentences to make it 5 sentences.
5. Try not to change the style of the sentence.

**Attention:**
1.Your answer only needs to give answer, no explanation is required!
2.Your answer should be in the same format as the example, given in JSON format

**input: {Input}**
**output: {Your Answer}**

Figure 6: Divide-and-conquer Prompt Template

## Task:

Extract entity and triples from a sentence (Before extracting triples, please fully understand the meaning of this passage and extract it after grasping the global semantics.)

**For example:**

***input:*** *{*
*"sentence": "A young man standing on stage wearing a white shirt and black pants."*
*}*
***output****:{*
*"entity": ["man","stage","shirt","pants"],*
*"attribute pairs": [*
*["man","young"],*
*["shirt","white"],*
*["pants","black"]*
*],*
*"triples": [*
*["man","stand","stage"],*
*["man","wear","shirt"],*
*["man","wear","pants"],*
*["man","is","young"],*
*["shirt","is","white"],*
*["pants","is","black"]*
*]*
*}*

**Attention:**

1.Your answer only needs to give answer, no explanation is required!
2.Strictly the same output format as the example, given in json format!
3.Your answer should be in strict JSON format and no comments should be added to the content!

**input: {Input}**
**output: {Your Answer}**

Figure 7: Scene-graph-generation Prompt Template