

A. Zoom-in Study regarding GPT-4V

We perform a zoom-in study of RLHF-V concerning GPT-4V to provide a better understanding of their behaviors.

A.1. Hallucination Patterns

We conduct a comparative analysis of the responses generated by RLHF-V and GPT-4V, and have the following key observations:

(1) Compared with RLHF-V, GPT-4V tends to describe more details in the images and elaborate more on the interrelations among them. Quantitatively, we utilize ChatGPT to extract all the object mentions in the responses of GPT-4V, and find that the average number per response is 2.1 times larger than RLHF-V. We mainly attribute this to the higher resolution (7.6 times larger than RLHF-V) [1] and the more powerful LLM backbone [37].

(2) GPT-4V’s hallucinations are more concentrated in some responses. In HumanEval, the hallucination rates of GPT-4V on *Object* and *Position* are comparable to RLHF-V. However, in the comprehensive *ALL* metric, the hallucination rate is 17.3% lower than RLHF-V. To better understand the reasons behind this phenomenon, we conduct a thorough analysis of the evaluation results. We observe that different types of hallucinations in GPT-4V are often concentrated in a small subset of responses, while contributing to hallucination rates across multiple subcategories. Quantitatively, we sort the responses of each model by the hallucination count in descending order, and plot the curve of hallucination count ratio vs hallucination response ratio. From the results in Figure 4, we can see that the top 45.6% hallucinated responses of GPT-4V contribute to 75% hallucination counts. In comparison, the top 64.6% hallucinated responses of RLHF-V contribute to 75% hallucinations. We refer readers to Section B for more qualitative results.

A.2. Distillation against GPT-4V

Upon observing GPT-4V’s superior fine-grained image perception and text generation capabilities, an intuitive question is, will it be beneficial to distill GPT-4V capabilities through visual instruction tuning? To this end, we collect 1.2k visual instruction data about long-form image descriptions from GPT-4V. We then use the response generated by GPT-4V to fine-tune our model. We observe that the average number of object mentions in the model response significantly increases by 1.8 times compared with the origin model. However, this can be a double-edged sword: as shown in Table 4, the hallucination rate significantly increases as well.

The results are consistent with the hypothesis of [2]: “If we supervise the model against instruction data that far exceeds its own foundational capabilities, we are essentially teaching the model to hallucinate.” Specifically, our

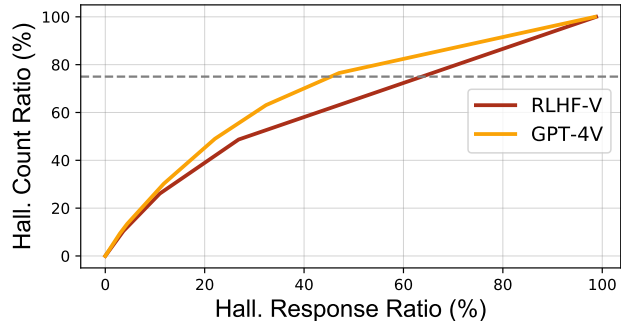


Figure 4. Distribution of hallucination segments over different responses. GPT-4V hallucinations are more concentrated on a smaller subset of the responses. Hall.: Hallucination.

model learns to produce more details and the interrelationship among them through distillation against GPT-4V, while the fundamental capabilities of the model are not enough for this demand. As a result, the hallucination problem is remarkably exacerbated. The results show that visual instruction data (or distillation target) is not the stronger the better, but rather should match the foundational capability of the model.

B. Qualitative Results

We provide more qualitative results in this section to facilitate a more intuitive understanding and comparison of different models. Based on the qualitative results, we have the following observations:

(1) RLHF-V typically exhibits less hallucination in both short-form QA and long-form QA scenarios, compared with open-source models such as LLaVA-RLHF and InstructBLIP, as shown in Figure 5, 6, 7, and 8.

(2) GPT-4V is more descriptive regarding details in images as shown in Figure 6, 8, 9 and 10. For example, in Figure 9, GPT-4V mentions *black dots* across each *tile* while RLHF-V does not describe these details.

(3) RLHF-V is more resistant to the over-generalization problem as shown in Figure 9 and Figure 10. In Figure 9, GPT-4V falsely mentions objects which are highly related to the scene while not shown in the image such as *exhaust*, *hood*, and *bottle*.

C. Implementation Details

We provide more implementation details in this section for better reproducibility. Benefiting from the high efficiency of training, we make all parameters trainable during the training process, which costs merely less than 1 hour on 8 A100 GPUs in total. We empirically find that adopting a longer warm-up (10% training steps) can make the training more stable and consequently apply this setting for all ex-

Model	HumanEval↓				MHB↓
	Obj.	Pos.	Num.	All	Resp.
Muffin [60]	33.6	16.4	26.0	74.7	68.8
RLHF-V	21.9	7.5	14.4	55.5	52.1
w/ GPT-4V distil.	45.2	10.3	20.6	75.3	62.5

Table 4. Experimental results of distillation against GPT-4V. MHB: MMHal-Bench, GPT-4V distil.: instruction-tune the model using responses generated by GPT-4V.

periments in this paper. As for data collection, besides the prompts obtained from [60], we also use image description prompts generated by GPT-4 during the annotation process which are listed in Table 5.

D. Evaluation Details

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

D.1. Baselines

We compare with a series of state-of-the-art baselines:

- **LLaVA**: LLaVA [35] constructs 150K multimodal instructions based on the COCO dataset by asking GPT-4 to generate multi-turn dialogues for each image.
- **Muffin**: Muffin [60] propose to reformulate pre-trained vision-language models as the bridge toward large language models. The model is firstly pre-trained on 180M image-text pairs and then fine-tuned on their proposed UniMM-Chat instruction dataset consisting of 1.1M multimodal instructions.
- **LRV**: LRV [33] is fine-tuned on 400K instruction data generated by GPT-4, and mitigates hallucination by limiting the response length.
- **LLaVA-RLHF**: The concurrent LLaVA-RLHF employs the strong 13B Vicuna 1.5 [62] (fine-tuned from LLaMA-2) as LLM backbone. It first trains the model with 122K instructions from VQAv2 [18], A-OKVQA [46] and Flickr30k [40] to improve the foundational capabilities of the model. It then trains the reward model on 10K human-labeled preference data, and performs proximal policy optimization [45] on 72K factually augmented data.
- **InstructBLIP**: InstructBLIP [14] constructs a multimodal instruction tuning dataset based on 26 public datasets by apply pre-defined templates to directly formulate these datasets into a unified format. They also devise a novel instruction-aware Q-Former and train the model on the proposed dataset.
- **Qwen-VL-Chat**: Qwen-VL-Chat [6] utilizes a large ViT with 1.9B parameters initialized from OpenCLIP’s bigG [21] as image encoder. It is pre-trained on 1.4B

image-text pairs and fine-tuned on more than 50M high-quality multimodal instructions.

- **LLaVA 1.5**: LLaVA 1.5 [34] also employs the strong 13B Vicuna 1.5 [62] (fine-tuned from LLaMA-2) as LLM backbone. It is pre-trained on 558K selected image-text pairs and fine-tuned on 665K multimodal instructions with elaborately designed training strategies.

D.2. Benchmarks

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

- **Object HalBench**: Object HalBench [43] is a widely adopted benchmark for assessing object hallucination in detailed image descriptions. To improve the evaluation stability, we augment the benchmark with 8 diverse prompts for detailed image descriptions during evaluation, where 4 instructions are adopted from [19] and the other 4 instructions are generated by GPT-4. We confirm that there is no overlap between the evaluation instructions and the training instructions. Detailed instructions are listed in Table 6. Following [29], we randomly sample 300 images from the validation set of COCO [31] to form the evaluation image set. Regarding metrics, the response-level hallucination rate is the number of responses with object hallucinations divided by the number of responses that introduce COCO objects, while the mention-level hallucination rate is the number of falsely mentioned COCO objects in the generated responses divided by the total number of mentioned COCO objects. During evaluation, we first generate descriptions on images from the benchmark and then leverage ChatGPT to extract the mentioned objects in these responses which are further used to calculate the final scores following. Unlike [43] which detects object mentions by exact-match, we find ChatGPT can perform the extraction with both better precision and recall and consequently apply this setting during evaluation. The full prompt we used to conduct such extraction is shown in Table 7.
- **MMHal-Bench**: MMHal-Bench [48] evaluates hallucinations and response informativeness. It consists of 96 images from the validation and test sets of OpenImages [23]. Each image in this benchmark is annotated with a brand new question and the image-question pairs cover 12 common object meta-categories from COCO.
- **HumanEval**: The above evaluations are either limited to common object hallucination or dominated by short-form question answering (i.e., questions that can be sufficiently answered by a few words). To provide a more reliable and comprehensive evaluation over diverse hallucination types, we present HumanEval benchmark, which covers both long-form image descriptions, and short-form questions. The benchmark contains 146 samples collected from Object HalBench (50) and MMHal-Bench

- Identify and describe each object in the image in detail.
- Describe the key features of the image in great detail.
- What are the main elements in this image? Describe them thoroughly.
- Explain what's happening in the image with as much detail as possible.
- Detail the image's components with particular focus on each entity.
- Provide an intricate description of every entity in the image.
- What are the main objects or subjects in the image? Please describe them in detail.
- What is the setting or environment in which the image takes place?
- How do the elements in the image relate to each other in terms of positioning or composition?
- Explain the elements of the image with thorough attention to detail.
- Explain the image's various components in depth.
- What are the key features you observe in the image?
- Can you point out the details that make this image unique?
- Itemize the elements you identify in the image and describe them thoroughly.
- Convey the specifics of the image with meticulous attention to detail.
- Tell me what catches your eye in the image, and describe those elements in depth.

Table 5. The list of instructions for detailed image description used in training.

- Provide a thorough description of the given image.
- What is this photo about? Please answer in great detail.
- Provide a thorough description of the given picture.
- Explain the narrative or story that the image seems to convey, detailing each part that contributes to it.
- Please provide a detailed description of the image. Describe the visual elements, colors, shapes, textures, and any objects or people present along with the overall mood or atmosphere portrayed in the image.
- Please provide a detailed description of the image, including its visual elements, such as colors, shapes, textures, objects, and people.
- Provide an intricate description of the image, capturing its visual elements, including colors, shapes, textures, objects, and any people present.
- Compose a detailed account of the image, encompassing its visual characteristics, like colors, shapes, textures, objects, and any human subjects, by paying careful attention to the specifics.

Table 6. The list of instructions for Object HalBench.

You are an expert in image objects extraction according to a question answer pair. We asked an examiner to answer a question about a picture.

[Start of Question]

<image> {question}

[End of Question]

[Start of Examiner's Answer]

{answer}

[End of Examiner's Answer]

Assume that the answer is correct, please identify all visible objects that are directly shown in the image. Please following the instructions in below:

1. You should only mention objects that are explicitly mentioned in the examiner's answer.
2. You should only extract the object names without the attributes of the objects.
3. You should not include the properties of the object, like the color, material, etc. as part of the object name in your result.
4. Make your answer precise. Present the results in a JSON list format: ["object_1", ..., "object_n"].
5. You should return an empty JSON list () if no visible objects can be found.

Table 7. The prompt we used to extract object mentions from image captions with ChatGPT.

(96). Given model responses, we ask human annotators to label the hallucinated segments and hallucination types of the segments, including objects, positions, numbers and others. We report the response-level hallucination rate on these types.

- **LLaVA Bench:** LLaVA Bench [35] is a widely adopted benchmark for assessing multimodal conversation, detailed description and complex reasoning capabilities. It consists of 30 image-question pairs for the aforementioned three capabilities respectively and scores model output against reference response via GPT-4.
- **VQAv2:** VQAv2 [18] is a dataset for short-form visual question answering. The test-dev set of VQAv2 consists of 107K image-question pairs which covers a diverse range of capabilities.



Short-form QA

What is the man doing in the image?

⊗ RLHF-V (This work)

In this picture, the man is **giving a thumbs up gesture**. The thumbs up sign is a universally recognized symbol of approval or agreement, often used to indicate satisfaction with something. It seems like he may be reacting positively to something that is happening around him.

⊗ InstructBLIP

The man in the image is sitting at a table, **talking on his cell phone**.

⊗ LLaVA-RLHF

The man in the image is **talking on his cell phone** while sitting in a crowded room. He is surrounded by other people, **some of whom are also sitting and talking on their phones**. The scene suggests that the man is engaged in a conversation, possibly discussing something important or casually chatting with someone on the other end of the line.

The presence of **multiple people using their cell phones** in the same room indicates that the image may have been taken in a public setting, such as a restaurant, a waiting area, or a social event. It is common for people to use their phones in such settings, as it allows them to stay connected with friends, family, or colleagues, or to attend to work-related matters.

The man's conversation could be a personal one, or it could be related to a professional matter, such as discussing a project or coordinating plans with a colleague. It is also possible that **he is using his phone to access information, such as checking emails, browsing the internet, or using social media**.

In any case, the image captures a moment in time where the man is actively engaged in a conversation on his cell phone, surrounded by **other people who are also using their devices to communicate or attend to their needs**. This scene reflects the ubiquity of cell phones in modern life, and the way they have become an integral part of our daily routines and social interactions.

⊗ GPT4-V

The man in the image appears to be **scratching or touching his head** with one hand while **giving a thumbs-up gesture** with the other hand.

Figure 5. Qualitative results of different models. **Correct answers** and **hallucinations** are highlighted in color respectively.



Long-form QA

What do you think is going on in this snapshot?

RLHF-V (This work)

In this image, we see a small gray elephant standing alone in a dirt field. The elephant appears to be the focal point of the photo, with its impressive size and unique color. In the background, there is a building that complements the natural surroundings very well.

InstructBLIP

The image depicts an elephant standing in a sandy area near a wall. The elephant is the main focus of the scene, **occupying most of the space in the frame**. There are several other objects scattered around the area, including **a few chairs placed near the elephant**. In addition to the elephant and **chairs**, there are **two tires** visible in the scene. **One tire is positioned closer to the left side of the image**, while the other is located on the right side. **Both tires appear to be partially buried in the sand**, possibly indicating that they have been placed there for the elephant to interact with. Overall, the scene provides a glimpse into the life of an elephant in a controlled environment.

LLaVA-RLHF

The image features a large elephant standing in a sandy area, surrounded by rocks and a building. The elephant is positioned near the center of the scene, with its trunk hanging down.

In the background, there are two tires lying on the ground, possibly indicating that the area is used for various activities or training exercises. Additionally, there is **a bird perched on a rock** in the scene, adding a touch of nature to the environment.

GPT4-V

The image depicts an elephant in an enclosure, likely at a zoo or a sanctuary. The environment has a concrete wall, a door, and some elements like rocks, a blue tarp, and a tire, possibly used as play or enrichment items for the elephant. The scene suggests that this is a space where the elephant is housed, and the items around might be used to keep it engaged or stimulated.

Figure 6. Qualitative results of different models. **Hallucinations** are highlighted in color.



Long-form QA



What might be the significance or role of the objects you described in ensuring city traffic flows smoothly?

⊗ RLHF-V (This work)

The objects in the image hold significance in ensuring city traffic flows smoothly as they help regulate and control the movement of vehicles and pedestrians. The traffic lights, for instance, provide visual cues to drivers and pedestrians, indicating when it is safe to proceed, stop, or yield. By doing so, they help reduce traffic congestion, prevent accidents, and improve overall safety.

Additionally, the street markings, such as the crosswalk and bike lane, provide clear guidance to drivers, cyclists, and pedestrians by outlining designated areas for each group. These markings contribute to a more organized and safer traffic flow, as they indicate the expected behavior of different users in sharing the road or navigating the intersection.

⊗ LLaVA-RLHF

The objects and features described in the image play a crucial role in ensuring that city traffic flows smoothly and safely. Here are some of the key aspects they contribute to:

1. Traffic lights: The presence of traffic lights at the intersection helps regulate the flow of traffic by controlling the right of way for vehicles and pedestrians. They ensure that traffic moves in an orderly manner, reducing the risk of accidents and congestion. By following the traffic light signals, drivers can safely navigate the intersection and avoid collisions with other vehicles or pedestrians.

2. **Parking meters:** **Parking meters** play a vital role in managing on-street parking in busy urban areas. They help maintain a steady flow of traffic by preventing vehicles from occupying parking spaces for extended periods, which can lead to congestion and reduced accessibility for other drivers. **Parking meters** also generate revenue for the city, which can be used to fund public services and infrastructure projects.

3. Signage: The presence of various signs on the street, such as **no parking signs**, helps communicate important information to drivers and pedestrians. **These signs provide guidance on parking restrictions**, speed limits, and other traffic rules, ensuring that everyone on the road is aware of their responsibilities and the potential consequences of violating traffic regulations.

4. Bike lanes: The presence of bike lanes in the image indicates that the city is promoting alternative modes of transportation, such as cycling. Bike lanes provide a dedicated space for cyclists, reducing the risk of accidents between bikes and cars. They also encourage more people to choose cycling as a means of transportation, which can help reduce traffic congestion and improve air quality in the city.

5. Pedestrian crossings: The presence of pedestrian crossings in the image highlights the importance of ensuring the safety of pedestrians in busy urban areas. These crossings provide designated areas for pedestrians to cross the street safely, reducing the risk of accidents and promoting a more walkable city environment.

In summary, the objects and features described in the image, such as traffic lights, **parking meters**, signage, bike lanes, and pedestrian crossings, play a significant role in ensuring that city traffic flows smoothly and safely. They help regulate traffic flow, manage parking, communicate important information to drivers and pedestrians, promote alternative modes of transportation, and prioritize the safety of pedestrians. By implementing these measures, cities can create a more efficient and sustainable urban environment for all residents.

Figure 7. Qualitative results of different models. **Unreasonable reasoning and extensions** and **scene related hallucinations** are highlighted in color respectively.



Long-form QA

What might be the significance or role of the objects you described in ensuring city traffic flows smoothly?

InstructBLIP

In the image, there are several objects that play a significant role in ensuring city traffic flows smoothly. These objects include streetlights, traffic lights, **cars, bicycles, and pedestrians**. Streetlights provide illumination for drivers and pedestrians during nighttime or low-light conditions, improving visibility and safety on the road. Traffic lights regulate the flow of traffic at intersections, ensuring that vehicles, bicycles, and pedestrians follow the rules of the road and avoid collisions. **Cars and bicycles are essential components of the transportation system, contributing to the movement of people and goods within the city. Pedestrians also play a crucial role in ensuring the smooth flow of traffic by crossing streets safely and adhering to traffic rules.** Overall, these objects work together to create a safe and efficient transportation system in the city.

GPT4-V

The objects in the image are various traffic control devices used in urban settings to manage and regulate the flow of traffic, ensuring safety and order on the roads. Here's their significance and role:

1. Traffic Lights: They control the flow of traffic at intersections, directing vehicles when to stop and go. This prevents collisions and ensures an orderly movement of traffic.
2. Street Signs: Various signs in the image, like "No Right Turn", inform drivers of local traffic laws and regulations they must follow, such as which turns are prohibited at an intersection.
3. Bicycle Lane Markings: The green painted bike lane signifies a dedicated space for cyclists, which helps in reducing conflicts between motor vehicles and bicycles.
4. Pedestrian Signals: Pedestrian traffic lights, although not clearly visible in this image, usually accompany traffic lights to indicate when it is safe for pedestrians to cross the street.
5. Road Markings: The various lines and symbols painted on the road surface guide drivers on how to use the lanes properly, where to stop, and the direction of permissible traffic movement.
6. Parking Signs: These signs inform drivers of the areas where parking is either allowed or prohibited during certain times, which contributes to the reduction of roadside congestion.

Each of these devices plays a crucial role in maintaining a systematic and smooth flow of different modes of transportation within the city streets, minimizing accidents, and ensuring the safety of all road users.

Figure 8. Qualitative results of different models (continued figure). **Unreasonable reasoning and extensions** are highlighted in color.



Long-form QA



Provide a thorough description of the given picture.

⚙️ RLHF-V (This work)

The image depicts a small kitchen with **white walls** and wood cabinets. There is a center island **in the middle of** the kitchen, providing additional counter space. The kitchen is equipped with standard appliances, such as a stove, located on the right side, and a sink on the left.

⚙️ GPT4-V

Certainly! The image showcases an interior view of a kitchen. Here are the details:

Cabinetry: The kitchen is equipped with a combination of **white** overhead and **base cabinets, some of which have a horizontal black line running along their middle**. The design appears modern, with a sleek finish.

Countertop: The countertop appears to be made of a light-colored, possibly marble or granite material. The main countertop extends along one side of the kitchen, while an additional rounded extension provides a potential breakfast or serving area.

Flooring: The floor is checkered with white tiles, and numerous black dots are scattered across each tile, creating a distinct pattern.

Appliances and Fixtures: On the right, there's a silver, freestanding gas range and oven. **Just above it is an exhaust or range hood**. A stainless-steel sink is set into the countertop on the left side of the image, **beneath which there's a dish drying rack attached to the wall**.

Lighting: The ceiling is equipped with multiple **recessed lights**, ensuring the kitchen is well-lit.

Other Details: A **green trash bin** is placed beside the gas range. Various other minor items, like **a bottle and some containers**, can be seen on the countertop. There's also a metallic vertical pole, possibly **a utility or storage pole, standing between the range and the rounded counter extension**.

Walls and Door: The kitchen walls are painted in a muted beige or light brown color. To the far right, there's a wooden door, possibly leading to a pantry or another room.

Ambiance: The kitchen exudes a clean, functional, and organized atmosphere, with a blend of neutral tones and contemporary design elements.

Overall, the kitchen appears to be well-maintained and offers a mix of practicality and style.

Figure 9. Qualitative results of different models. **Hallucinations** and *scene related hallucinations* are highlighted in color respectively.



Long-form QA



Please provide a detailed description of the image. Describe the visual elements, colors, shapes, textures, and any objects or people present along with the overall mood or atmosphere portrayed in the image.

RLHF-V (This work)

The living room is clean and empty of people, with the sun shining in through the **windows**. It features a **white rug** and black furniture that contrasts with the light from the **windows**. There is a doorway leading to a balcony that showcases the railing. The living room contains a couch, **a table**, and two chairs.

GPT4-V

The image showcases an elegantly designed living room.

Starting with the overall atmosphere, the room exudes a warm and inviting ambiance. Natural sunlight floods in from large **windows on the left**, emphasizing the spaciousness of the room and creating a luminous glow.

As for the visual elements:

Colors: The room primarily features neutral tones—creams, browns, and blacks. These muted colors contribute to the room's calming and sophisticated feel. The wooden floor has a polished honey-brown hue which complements the lighter walls.

Furniture: A variety of seating options are available. There are two dark leather chairs **on the left side of the room, facing the window**. In the center, a black leather couch with decorative cushions sits opposite a glass coffee table with metallic legs. **On the far right, there's another seating area with two plush chairs flanking a fireplace.**

Decorative Elements: Above the fireplace is a white mantel adorned with decorative items. On the left wall, a sleek table holds decorative items like **a lamp and a vase** with flowers. **Above this table, a large mirror is hung**, reflecting the room's light and enhancing its spaciousness. **On the adjacent wall**, three framed artworks are neatly aligned, adding an artistic touch to the room.

Textures: The room presents a blend of textures. The plushness of the rug contrasts with the smooth leather of the furniture, the glass surface of the coffee table, and the sleek wooden floor.

Shapes: There are various shapes at play. The rectangular and square forms of the artworks, the **mirror**, and the **windows** are balanced by the **rounded shapes of the seating and the coffee table.**

Plants: A large potted plant sits by the window, adding a touch of nature and a refreshing green to the neutral palette.

The room is a blend of modern and classic elements, providing a sense of comfort while maintaining a refined aesthetic.

Figure 10. Qualitative results of different models. **Hallucinations** and *scene related hallucinations* are highlighted in color respectively.