

# Supplementary Material For INTER: Mitigating Hallucination in Large Vision-Language Models by Interaction Guidance Sampling

## 1. Details of the Benchmarks

**POPE.** The Polling-based Object Probing Evaluation (POPE) [15] utilizes images sampled from several datasets, including MSCOCO [17], A-OKVQA [22], and GQA [12]. Every question in POPE is *"Is there a <object> in the image?"*. For each dataset, it incorporates random, popular, and adversarial question sampling strategies to sample <object> and create three partitions. Random represents randomly selecting an object from the candidate object set. Popular means selecting the objects that occur more frequently. Adversarial refers to select objects that have a high co-occurrence frequency with the objects in the image. Therefore, the adversarial partition is the most challenging, as hallucinations are often caused by a high co-occurrence frequency between objects.

**MME.** MME [8] evaluates LVLMs using 14 subtasks from the perspectives of perception and cognition. There are four subtasks for the evaluation of the cognition ability, including commonsense reasoning, numerical calculation, text translation, and code reasoning. The remaining subtasks are used to evaluate perceptual abilities from the perspectives of coarse grained recognition, fine grained recognition, and OCR. Each image corresponds to two questions with opposing answers. For each subtask, the score of LVLMs is represented by the proportion of all questions answered correctly, as well as the proportion of both questions for each image answered correctly.

**MM-Bench.** MM-Bench [20] employs 20 subtasks to evaluate LVLMs in detail. These 20 subtasks are further divided into six perspectives: 'Coarse Perception (CP)', 'Cross-instance Fine-grained Perception (FP-C)', 'Single-instance Fine-grained Perception (FP-S)', 'Attribute Reasoning (AR)', 'Logic Reasoning (LR)', and 'Relation Reasoning (RR)'. For each sample, MM-Bench sets several options and requires the LVLMs to return one of them. The template for each question is *'Answer with the option's letter from the given choices directly.'* More importantly, MM-Bench creates questions with the same content but differing option sequences by repeatedly rotating the order of them. Then, for each sample, the accuracy across all orders is collected, and if all are answered correctly, the LVLMs score for that sample. Therefore, MM-Bench's evaluation

of LVLMs is more rigorous and is not influenced by the order of the options.

**MMStar.** Like MM-Bench, MMStar [3] also establishes multiple subtasks and categorizes them into six perspectives: 'Coarse Perception (CP)', 'Fine-Grained Perception (FP)', 'Instance Reasoning (IR)', 'Logical Reasoning (LR)', 'Science & Technology (ST)' and 'Math (MA)'. Every aspect have three subtasks. But the difference is that MMStar uses a four-tier filtering mechanism to select 1,500 elite samples from an initial pool of 22,401 samples. Each sample strictly adheres to three criteria during the filtering process: it must rely on visual content comprehension, cover a broad range of ability dimensions, and require advanced multimodal reasoning capabilities. Therefore, using MMStar for evaluation can better reflect the capabilities of LVLMs.

**CHAIR.** CHAIR [21] has established two metrics,  $CHAIR_S$  and  $CHAIR_I$ , to assess the degree of hallucination in the generated responses. Where  $CHAIR_S = \frac{|\{captions\ with\ hallucinated\ objects\}|}{|\{all\ captions\}|}$  indicates the degree of hallucination at the sentence level, while  $CHAIR_I = \frac{|\{hallucinated\ objects\}|}{|\{all\ mentioned\ objects\}|}$  represents the degree of hallucination at the object level. Following previous work, we randomly sampled 500 samples and used *'Please describe this image in detail.'* to guide the LVLMs in generating captions for the images.

## 2. Result on InternVL2.5-MPO

In order to further demonstrate the effectiveness of INTER, we conducted a comparison on the current state-of-the-art LVLm InternVL2.5-MPO (8B) [4]. As shown in Tab. 1, the performance of INTER is superior to the baseline methods across various benchmarks. Moreover, 'Nucleus+INTER' performs better than VCD [6] across all benchmarks, while 'Beam+INTER' also performs better than OPERA [11].

## 3. Ablation Study on Interaction Guide Locator.

In addition to the effectiveness analysis of the Interaction Guide Locator based on Beam Search [2, 9, 24], we also conducted ablation experiments on various decoding strate-

model	benchmark	Nucleus	Nucleus+INTER	Beam	Beam+INTER	VCD*	VCD*+INTER	OPERA <sup>†</sup>
InternVL2.5-MPO (8B) [19]	MME (Total Score) [8] ↑	2175.7	<b>2204.8</b>	2298.3	<b>2316.4</b>	2189.2	<b>2209.9</b>	2299.7
	POPE (MSCOCO) [16] ↑	85.7	<b>89.2</b>	88.7	<b>89.3</b>	88.6	88.5	88.9
	MM-Bench [20] ↑	80.1	<b>81.5</b>	84.4	<b>84.6</b>	80.8	<b>81.6</b>	84.4
	MMStar [3] ↑	60.8	<b>62.5</b>	63.0	<b>63.9</b>	61.9	<b>63.3</b>	63.5
	CHAIR ( $C_S+C_I$ ) [21] ↓	25.2	<b>21.6</b>	23.6	<b>19.7</b>	25.5	25.9	22.0
	LLaVA-Bench [18] ↑	9.5	<b>11.9</b>	9.3	<b>12.5</b>	10.1	<b>11.2</b>	10.5

Table 1. **Validation of INTER on the state-of-the-art LVLM InternVL2.5-MPO [4].** \* and <sup>†</sup> represent correction based on Nucleus Sampling and Beam Search.

method	InstructBLIP [5]	LLaVA-v1.5 [19]	mPLUG-owl2 [29]
○ Nucleus+IPM	1569.7	1690.9	1640.6
● Nucleus+INTER	<b>1595.5</b>	<b>1731.6</b>	<b>1641.7</b>
○ Beam+IPM	1556.2	1648.6	1623.0
● Beam+INTER	<b>1562.2</b>	<b>1744.0</b>	<b>1716.1</b>
○ VCD*+IPM	1583.6	1700.0	1620.1
● VCD*+INTER	<b>1605.0</b>	<b>1749.6</b>	<b>1626.2</b>
○ OPERA <sup>†</sup> +IPM	1553.5	1720.8	1625.7
● OPERA <sup>†</sup> +INTER	<b>1567.0</b>	<b>1727.4</b>	<b>1741.7</b>

Table 2. **Ablation Study on Interaction Guide Locator (IGL).**

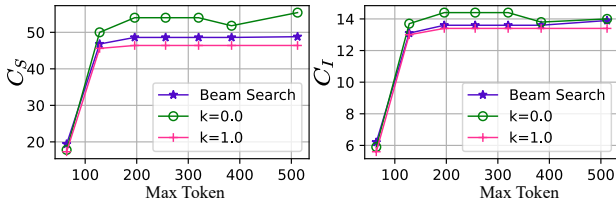


Figure 1. **Parameter analysis of  $k$  in Interaction Guide Locator.** Evaluation of  $C_I$  and  $C_S$  after using different  $k$  to guide Beam Search [2, 9, 24] on various lengths on CHAIR [21].

gies for IGL. As shown in Tab. 2, we evaluated the performance improvement brought by IGL on MME [8]. It can be observed that the performance significantly decreases without IGL across all decoding strategies, suggesting that IGL identifies the positions of keywords, preventing the excessive guidance of interactions, thereby effectively improving performance.

#### 4. Parameter Analysis of Interaction Guide Locator.

Through experiments on CHAIR [21] and MME [8] benchmarks, we analyze how the interaction guidance coefficient  $k$  affects the performance of INTER.

As shown in Fig. 1, varying  $k$  values lead to significantly different behaviors in LLaVA-v1.5. When  $k = 0.0$  which applies the Interaction Probability Modifier at all decoding steps, we observe reduced hallucination for short sequences after using INTER. However, this approach harms perfor-

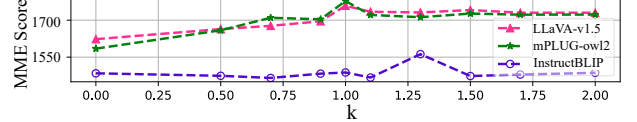


Figure 2. **Parameter analysis of  $k$  on MME [8].** Each value represents total score of using INTER on Beam Search [2, 9, 24].

method	InstructBLIP [5]	LLaVA-v1.5 (7B) [19]	Qwen-VL [1]	mPLUG-owl2 [29]
○ Nucleus [10]	77.0	79.1	76.1	76.8
● Nucleus+INTER	<b>81.9</b>	<b>84.3</b>	<b>81.9</b>	<b>80.2</b>
○ Greedy [23]	81.3	85.1	79.4	80.9
● Greedy+INTER	<b>82.2</b>	<b>85.2</b>	<b>81.4</b>	<b>80.9</b>
○ Beam [2, 9, 24]	81.4	84.7	79.7	80.2
● Beam+INTER	<b>83.3</b>	<b>84.7</b>	<b>81.2</b>	<b>81.0</b>
○ VCD* [6]	80.6	82.7	82.3	79.7
● VCD*+INTER	<b>80.9</b>	<b>83.6</b>	<b>82.0</b>	<b>79.5</b>
○ OPERA <sup>†</sup> [11]	81.3	84.9	79.8	80.4
● OPERA <sup>†</sup> +INTER	<b>82.5</b>	<b>85.8</b>	<b>83.1</b>	<b>81.1</b>

Table 3. **Evaluating the performance of INTER’s correction on four decoding strategies** by the mean F1-score across various partitions of GQA [12]. Higher values are better.

mance in longer sequences due to unnecessary modifications at non-critical positions, as evidenced by the performance drop compared to  $k = 1.0$ .

Fig. 2 reveals model-dependent optimal  $k$  values. On MME, InstructBLIP achieves peak performance at  $k = 1.3$ , beyond which excessive adjustment suppression causes gradual performance degradation. This suggests a balance between necessary corrections and interference avoidance.

#### 5. Result on POPE

In this subsection, we evaluate the performance of the proposed INTER on the GQA [12] dataset within the POPE benchmark. The results, as shown in the Tab. 3, indicate that significant performance improvements across four models. Furthermore, these enhancements are consistent with the results on the MSCOCO [17] and AOKVQA [22] datasets, further validating the effectiveness and robustness of our

approach.

## 6. Result on MME

In addition to demonstrating the performance improvements brought by INTER across various decoding strategies in 14 subtasks, we also conducted comparisons in terms of the total score and perception total score of MME [8]. As shown in Tab. 4, after correction with INTER, there was a maximum increase of over 343.7 points in the total score compared to Nucleus Sampling, and a maximum increase of over 311.2 points in the perception total score. Furthermore, it can be observed that there is a certain degree of improvement across different models and decoding strategies, indicating the effectiveness of INTER.

## 7. Result on MM-Bench

To illustrate the improvement of INTER on MM-Bench in more detail, we present the performance of each subtask in Tabs. 5, 14 and 15. As we can see, using INTER results in an improvement across various metrics. In addition, to validate the performance of INTER across different LVLMs, Tab. 6 presents the performance on mPLUG-owl2. It can be observed that there is a high consistency with LLaVA-v1.5, and INTER brings a certain degree of enhancement. Finally, detailed results of mPLUG-owl2 at each subtasks are also presented in Tabs. 16 and 17.

## 8. Result on MMStar

Likewise, to assess the effectiveness of INTER on MMStar, we also present the performance of each subtask on LLaVA-v1.5 (7B) [19] in Tabs. 7, 18 and 19. The results indicate that our approach achieves good performance across most subtasks. Although there is no improvement of the correction effects on VCD [6] and OPERA [11] in the ‘Math’, the correction results using INTER for ‘Nucleus’ outperform those of VCD, and the performance on Beam Search is better than OPERA.

In addition, we conducted comparative experiments on MMStar using mPLUG-owl2 in Tabs. 8, 20 and 21, and the results show that our method has a certain corrective effect across different LVLMs.

## 9. Result on Greedy Search

In Tabs. 3 to 8, 10, 11 and 14 to 21, we demonstrated the effectiveness of INTER in correcting the Greedy Search across various benchmarks. It is evident that there is a significant improvement across different benchmarks, indicating that our method INTER exhibits generalization capabilities in correcting various decoding strategies.

## 10. Result on LLaVA-Bench

To more intuitively demonstrate the performance of INTER, detailed case studies were conducted using LLaVA-Bench. In Figs. 5 to 7, examples of the captioning and complex reasoning task for each model are presented. The hallucination parts are highlighted in red.

In addition to case study, we also assessed the accuracy and detailedness of responses generated by various methods on the LLaVA-Bench dataset using GPT-4o [14]. As shown in the Fig. 4, the answers generated after applying INTER calibration received higher scores. The template of prompt is shown in Fig. 3.

## 11. Computation Efficiency

Similar to VCD [6], which require additional forward passes, INTER also necessitates extra inference to compute the logits under different subsets of  $A$ . While INTER increases the total number of forward passes, the actual runtime overhead remains negligible due to the capability of compressing all subset evaluations into a single batch.

## 12. Comparison with Other Methods.

We conducted experiments with M3ID [7], Ritual [26] and SID [13] in Tab. 9. The results demonstrate that our INTER achieves comparable performance among compared methods.

## 13. Performance on Other Types of Tasks or Different LVLMs.

We conducted experiments with DeepSeek-VL2 [27] on the visual grounding task. As shown in Tab. 12, results show that the INTER boosts the model performance on this task.

## 14. The Range of the Harsanyi dividend.

The value range of  $I(A)^{y_t}$  could be influenced by several factors, *e.g.*, benchmarks, LVLMs, etc. These complexities make it challenging to establish a theoretical bound for its value range. Nevertheless, we conducted experiments to empirically assess the distribution of  $I(A)^{y_t}$  in Tab. 13. Moreover, when  $I(A)^{y_t}$  is negative, we consider that such interaction effects may hinder sampling this candidate token, which is considered similarly in prior studies [25, 31].

## References

- [1] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 2023. 2, 4, 9, 10

method	InstructBLIP [5]		LLaVA-v1.5 (7B) [19]		Qwen-VL [1]		mPLUG-owl2 [29]	
	Perception Total	Total	Perception Total	Total	Perception Total	Total	Perception Total	Total
○ Nucleus [10]	984.4	1251.8	1279.2	1502.2	1216.6	1465.6	1266.3	1573.5
● Nucleus+INTER	1295.6 (↑311.2)	1595.5 (↑343.7)	1372.0 (↑92.8)	1731.6 (↑229.4)	1279.8 (↑63.2)	1542.9 (↑77.3)	1353.8 (↑87.5)	1641.7 (↑68.2)
○ Greedy [23]	1160.9	1419.8	1452.2	1750.4	1238.9	1512.5	1352.5	1709.3
● Greedy+INTER	1291.2 (↑130.3)	1593.3 (↑173.5)	1470.5 (↑18.3)	1761.3 (↑10.9)	1292.4 (↑53.5)	1544.2 (↑31.7)	1360.4 (↑7.9)	1731.4 (↑22.1)
○ Beam [2, 9, 24]	1128.9	1318.6	1409.4	1707.2	1229.5	1513.4	1358.4	1710.5
● Beam+INTER	1281.8 (↑152.9)	1562.2 (↑243.6)	1438.3 (↑28.9)	1744.0 (↑36.8)	1271.5 (↑42.0)	1575.0 (↑61.6)	1363.3 (↑4.9)	1716.1 (↑5.6)
○ VCD* [6]	1167.9	1487.1	1364.0	1716.0	1240.0	1546.5	1269.7	1573.0
● VCD*+INTER	1306.8 (↑138.9)	1605.0 (↑117.9)	1380.0 (↑16.0)	1749.6 (↑33.6)	1297.0 (↑57.0)	1575.6 (↑29.1)	1305.2 (↑35.5)	1626.2 (↑53.2)
○ OPERA <sup>†</sup> [11]	1137.5	1326.5	1430.8	1721.2	1228.8	1501.7	1357.6	1740.8
● OPERA <sup>†</sup> +INTER	1274.1 (↑136.6)	1567.0 (↑240.5)	1439.8 (↑9.0)	1727.4 (↑6.2)	1304.0 (↑75.2)	1564.4 (↑62.7)	1377.7 (↑20.1)	1741.7 (↑0.9)

Table 4. **The total scores and perceptual total scores on MME [8].** ↑ means that higher values indicate lower hallucination levels. Results showed that the addition of INTER led to a certain degree of mitigating hallucinations for all decoding strategies.

#### GPT-4o Prompt

You are an AI tasked with evaluating and scoring the performance of two AI assistants in describing a specified image. Your evaluation will primarily focus on accuracy and detail in their descriptions. Accuracy will be assessed by identifying any hallucinations—elements of the description that do not align with the image and the related question. For detail, consider how comprehensive and rich the response is, excluding any hallucinated content. You will score each assistant on a scale from 1 to 10 based on these criteria. After scoring, you will provide an unbiased explanation of your evaluations, ensuring that your analysis is not influenced by the order in which the responses are presented.

#### Input format:

[Assistant 1]  
{Response 1}  
[End of Assistant 1]  
[Assistant 2]  
{Response 2}  
[End of Assistant 2]

#### Output format:

Accuracy: <Scores of the two responses>  
Reason:  
Detailedness: <Scores of the two responses>  
Reason:

Figure 3. Prompts of GPT-4o [14] for evaluations.

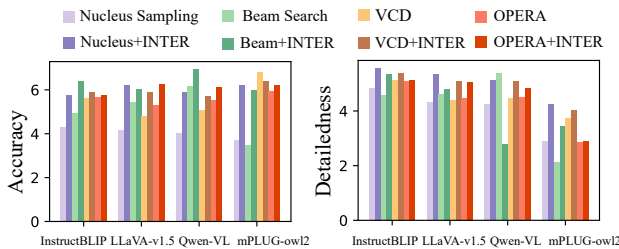


Figure 4. **Evaluating accuracy and detailedness on response of 60 Image-Text pairs in LLaVA-Bench [18] using GPT-4o [14].**

[2] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pas-

cal Vincent. Audio chord recognition with recurrent neural networks. In *ISMIR*, pages 335–340. Curitiba, 2013. 1, 2, 4, 8, 10, 11, 12

[3] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, et al. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*, 2024. 1, 2, 8, 11, 12

[4] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 1, 2

[5] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat





**Prompt:** Explain all the cues that indicate the current traffic conditions.

**Category:** complex

**Nucleus:** The image captures a busy city highway with heavy traffic, featuring numerous cars and trucks. The traffic lights control the flow of vehicles, which can be seen in various positions along the highway, both near and far. There are at **least 20 traffic lights visible in the image**, illustrating the high volume of vehicles and the well-maintained traffic control system. The presence of traffic lights along the highway suggests that the area is likely heavily congested and that drivers need to follow the traffic lights to prevent accidents and maintain order on the road. Furthermore, the nighttime setting implies that the traffic is active during typical rush hours, adding to the overall impression of considerable traffic congestion in the area.

**VCD:** In the image, the current traffic conditions are indicated by the numerous cars and trucks moving down the freeway at night. The vehicles are packed tightly together, with some cars and trucks visible in the foreground and others further back, creating a dense traffic situation. The darkness of the night adds to the feeling of congestion, as it is more difficult for drivers to see each other and maintain safe distances between their vehicles. The **traffic lights** also help guide the flow of traffic, but the heavy volume of vehicles and limited visibility make this nighttime commute a challenging one for drivers.

**OPERA:** In the image, the traffic conditions are quite congested, as evidenced by the numerous cars and trucks on the road. The vehicles are lined up in both directions, with some cars and trucks occupying the left side of the road, while others are on the right side. The congestion is further emphasized by the presence of multiple traffic lights, indicating that this is an urban area with a well-regulated traffic system. **The traffic lights** are scattered throughout the scene, with some located closer to the foreground and others further in the background. The combination of the large number of vehicles and the presence of traffic lights suggests that this is a busy and congested city street at night.

**INTER:** In the image, the cues that indicate the current traffic conditions include a large number of cars and trucks traveling on a busy highway at night. The vehicles are moving in both directions, creating a sense of congestion and slow-moving traffic. The presence of multiple cars and trucks, as well as the nighttime setting, suggest that this is a typical rush hour or peak travel time when people are commuting to and from work, school, or other destinations. The heavy traffic can lead to longer travel times, increased fuel consumption, and higher stress levels for drivers and passengers alike.



**Prompt:** Describe this photo in detail.

**Category:** detail

**Nucleus:** The image displays an open refrigerator with a variety of food items stored inside, including several milk bottles. There are **five bottles in total**, arranged throughout the fridge in different sections, suggesting a well-stocked and organized kitchen space. Additionally, there is an open box on a shelf displaying many different fruits. There are carrots inside the box, **as well as a group of oranges**. A bowl is located close to the left side, perhaps holding more fruits. There are two cups visible in the bottom right corner, placed separately, and possibly **used for consuming the food items**.

**VCD:** The image displays an open refrigerator filled with various food items. The refrigerator is packed with fresh fruits and vegetables, including **carrots, strawberries, apples, and bananas**. The carrots are arranged on the left side, with strawberries occupying the right side of the refrigerator. In addition to the fruits and vegetables, the refrigerator also contains dairy products such as a milk carton and **a juice box**. These items are scattered throughout the refrigerator, with the juice box placed on the left side and the milk carton closer to the center. The open door allows us to view the well-stocked and organized interior of the refrigerator.

**OPERA:** The image displays an open refrigerator filled with a variety of food items. The refrigerator is packed with numerous bottles, some of which are placed on **the top shelves** and others on the bottom shelves. In addition to the bottles, there are several carrots scattered throughout the refrigerator, with some located on the top shelves and others on the bottom shelves. A bowl can also be seen in the refrigerator, placed on one of the top shelves. The refrigerator appears to be well-stocked and organized, making it easy to find and access the food stored in the appliance.

**INTER:** The image features a refrigerator filled with a variety of food items. The refrigerator is packed with multiple containers, bottles, and other food items, creating a full view of the contents inside. There are several bottles placed in different areas of the refrigerator, with one on the left side, one near the top right corner, and another one positioned towards the middle-right section of the fridge. A bowl is placed in the lower part of the refrigerator, towards the left side. Additionally, the refrigerator contains a pile of carrots. The carrots are situated on the left side of the refrigerator, occupying the lower portion.

Figure 5. Case study on LLaVA-v1.5 through LLaVA-Bench. Hallucinations are marked in red.



**Prompt:** In what other ways might someone express the same sentiment that this meme is expressing?

**Category:** complex

**Nucleus:** The scene of the man and woman standing on a boat while hailing at the camera symbolizes a vulnerable moment in the context of the movie. In the unreleased Pirates of the Caribbean sequel, this particular image captures the genuine emotion and vulnerability as the characters **Jack and Lois** are separated from the rest of their companions, stranded in the midst of a turbulent ocean. This creates tension and anticipation as the audience waits to see how the characters will handle this perilous situation, emphasizing the challenges they will face while attempting to find their way back to their destination.

**VCD:** The scene of a man and a woman standing on a boat and **holding onto each other with their arms around each other** is significant in the context of the movie because it is a romantic moment. It signifies the bond and connection between the two characters and the development of their relationship. This moment can be seen as a turning point in their story, where their love and trust in each other become stronger, leading to a deeper connection. In movies, such scenes are often used to create an emotional connection with the audience and to emphasize the importance of love and companionship in the characters' lives.

**OPERA:** In the context of the movie "Titanic," this scene is significant because it captures the romantic moment between the two main characters, Jack and Rose, as they share their first kiss on the back of the boat. This scene is iconic and memorable, as it symbolizes the beginning of their love story and sets the tone for the rest of the movie. The image of Jack and Rose embracing each other on the boat, with the sunset in the background, evokes a sense of romance, adventure, and the enduring power of love.

**INTER:** In the context of the movie "Titanic," this scene is significant because it captures the romantic moment between the two main characters, Jack and Rose, as they share their first kiss on the ship's deck. This scene is iconic and memorable, as it represents the beginning of their love story and sets the tone for the rest of the movie. It also highlights the beauty of the ship and the breathtaking view of the ocean, which adds to the overall atmosphere and emotional impact of the scene.



**Prompt:** Please explain the meme in detail.

**Category:** detail

**Nucleus:** The image portrays a pug dog laying comfortably on the floor on a blue wood floor. The dog appears to be asleep or just resting. The scene is edited to include a funny-looking remark, stating that the dog is "just Monday." This phrase is displayed in the bottom part of the image and appears in a speech bubble, **giving an amusing touch to the scene. The size and placement of the dog are proportionately respected with the writing.**

**VCD:** The image shows a small, cute dog laying on its side on a wooden floor, seemingly exhausted or napping. The dog appears to be a puppy, and it is positioned on a wooden floor that is part of a deck. The scene is set against a light blue background. Below the dog, the words "Just Monday..." are displayed, adding a **playful and humorous touch to the scene.** The combination of the dog's adorable appearance and the light-hearted commentary creates a **fun and entertaining meme.**

**OPERA:** The image features a small brown dog lying down on a wooden floor, possibly a porch or a deck. The dog appears to be sleeping or resting comfortably. The scene is accompanied by a caption that reads "Monday. Just Monday." This meme is likely meant to convey a **sense of relaxation or a break** from the hustle and bustle of everyday life, as the dog seems to be enjoying its time on the floor.

**INTER:** The image features a small brown dog lying down on a wooden floor. The dog appears to be sleeping or resting, taking up a significant portion of the floor. Above the dog, there is a caption that reads, "Monday. Just... Monday." This meme is likely meant to convey the feeling of being tired or unenthusiastic about the start of a new week.

Figure 6. Case study on mPLUG-owl2 through LLaVA-Bench. Hallucinations are marked in red.



**Prompt:** Show me a detailed recipe for cooking this at home.

**Category:** complex

**Nucleus:** The image features a delicious sandwich cut into two equal halves to serve as subs. A green Subway logo is placed over the top of one half, emphasizing its originality and promoting their signature substyle dishes available there in this particular menu item called "Steak & Cheese!". In addition to these famous deli meats like ham or maybe salami inside each slice's bread roll with lettuce on both sides before adding slices from steaming grilled chicken breast!

**VCD:** The image features two cut subway sandwiches placed side by side on a dark background. Each sub is loaded with meats, lettuce, and possibly other vegetables, making it an appetizing dish to prepare at home. Subway's footlong sandwiches are popular choices for customers who crave a filling and tasty lunch option. These two subway sandwiches can be replicated easily using the ingredients you have available in your kitchen, ensuring that you can enjoy this satisfying meal anytime you want.

**OPERA:** The image features two sub sandwiches cut in half, placed on a black background. The sub sandwiches are loaded with various ingredients, including meat, lettuce, tomato, and cheese. The sandwiches are positioned close to each other, creating a visually appealing display. The sandwiches are likely from Subway, a popular fast-food chain known for its variety of sandwiches and fresh ingredients.

**INTER:** The sandwich is presented on a dark background, adding to the dramatic effect of the dish. The image captures the attention of anyone who appreciates high-quality, mouth-watering sandwiches. In order to recreate this dish at home, follow the steps below: 1. Start with a freshly baked or toasted sub roll. 2. Spread mayonnaise on the inside of the roll. 3. Layer on the desired fillings, such as meats, cheese, and lettuce. 4. Close the sandwich by placing the halves together, sealing the fillings in. 5. Enjoy your homemade, mouth-watering Subway-style sandwich!



**Prompt:** Please explain the meme in detail.

**Category:** detail

**Nucleus:** The image features a painting of an artistic portrayal, where dogs have been incorporated into the portrait. A dog can be seen standing in front and center on his two hind legs behind Da Vinci's Mona Lisa background scene-appearing as if he is her pet or co-star! In this clever composition by artist Chloe Jeavons Gardner based off one known most famous paintings from history "Mona", she added small details to pay homage both historic beauty while using contemporary touches like adding doggies at bottom right corner which makes it even more charismatic piece for sure!

**VCD:** The image features a dog dressed as Leonardo da Vinci's Mona Lisa. The dog is wearing a long, brown dress and appears to be looking directly at the viewer with its eyes. It's placed in front of an oil painting backdrop that resembles the famous artwork. In total, there are five dogs visible in the scene. Three of them can be found near the edges, while one sits on top of the other two towards the middle of the image.

**OPERA:** The image features a painting of a dog dressed up as Mona Lisa, the famous painting by Leonardo da Vinci. The dog is depicted in the style of the original Mona Lisa painting, with its head positioned towards the right side of the image and its body slightly tilted to the left. The painting captures the essence of the original Mona Lisa, while adding a unique twist by featuring a dog as the subject.

**INTER:** The image features a large, long-haired dog dressed up as Leonardo da Vinci's famous painting, "Mona Lisa". The dog, likely a dachshund, has been transformed into an artwork, complete with its own crown. It's clear that the artist put a lot of effort into creating this unique and intriguing piece.

Figure 7. Case study on InstructBLIP through LLaVA-Bench. Hallucinations are marked in red.



method	Overall	AR	CP	FP-C	FP-S	LR	RR
○ Nucleus [10]	57.3	50.0	73.0	48.8	58.1	44.7	52.6
● Nucleus+INTER	<b>62.6</b>	<b>57.0</b>	<b>76.4</b>	<b>52.0</b>	<b>65.0</b>	<b>48.3</b>	<b>63.2</b>
○ Greedy [23]	65.2	58.4	77.8	55.3	69.3	50.7	64.7
● Greedy+INTER	<b>65.2</b>	<b>58.4</b>	<b>77.8</b>	<b>55.3</b>	<b>69.3</b>	<b>50.7</b>	<b>64.7</b>
○ Beam [2, 9, 24]	65.1	57.8	77.8	55.3	69.3	50.7	64.7
● Beam+INTER	<b>65.1</b>	<b>58.1</b>	<b>77.8</b>	<b>55.3</b>	<b>69.3</b>	<b>50.7</b>	<b>64.7</b>
○ VCD* [6]	62.5	54.7	77.5	53.4	64.9	48.5	60.5
● VCD*+INTER	<b>62.9</b>	<b>54.7</b>	<b>77.6</b>	52.3	<b>66.0</b>	47.6	<b>65.0</b>
○ OPERA <sup>†</sup> [11]	65.0	57.8	77.8	55.3	69.0	50.7	64.7
● OPERA <sup>†</sup> +INTER	<b>65.0</b>	<b>57.8</b>	<b>77.9</b>	55.1	<b>69.0</b>	<b>50.7</b>	<b>64.7</b>

Table 5. Validation of INTER on MM-Bench [20] using LLaVA-v1.5 (7B) [19].

method	Overall	AR	CP	FP-C	FP-S	LR	RR
○ Nucleus [10]	57.0	51.1	72.2	41.5	59.9	44.6	55.3
● Nucleus+INTER	<b>61.4</b>	<b>57.6</b>	<b>74.4</b>	<b>41.7</b>	<b>62.8</b>	<b>46.9</b>	<b>60.8</b>
○ Greedy [23]	63.5	55.4	77.5	52.2	65.8	49.1	70.7
● Greedy+INTER	62.3	<b>56.3</b>	75.7	49.3	65.3	<b>49.3</b>	64.7
○ Beam [2, 9, 24]	63.5	55.4	77.5	52.2	65.8	49.8	69.2
● Beam+INTER	<b>63.7</b>	<b>55.6</b>	<b>77.5</b>	<b>52.4</b>	<b>65.9</b>	<b>49.8</b>	<b>70.7</b>
○ VCD* [6]	59.2	52.1	75.1	43.7	62.1	45.9	59.0
● VCD*+INTER	<b>59.5</b>	<b>53.6</b>	74.1	<b>44.3</b>	<b>62.4</b>	<b>48.0</b>	58.3
○ OPERA <sup>†</sup> [11]	63.4	55.4	77.5	52.2	65.8	49.1	69.2
● OPERA <sup>†</sup> +INTER	<b>63.6</b>	<b>55.4</b>	<b>77.5</b>	<b>52.3</b>	65.7	<b>49.8</b>	<b>70.9</b>

Table 6. Validation of INTER on MM-Bench [20] using mPLUG-owl2 [29].

method	Avg.	CP	FP	IR	LR	ST	MA
○ Nucleus [10]	29.3	52.8	22.4	38.0	22.8	17.6	22.4
● Nucleus+INTER	<b>31.9</b>	<b>58.0</b>	<b>29.8</b>	<b>39.4</b>	<b>27.6</b>	15.4	<b>22.8</b>
○ Greedy [23]	30.7	59.2	24.8	40.0	27.2	13.6	19.6
● Greedy+INTER	<b>31.9</b>	55.2	<b>29.2</b>	<b>45.2</b>	<b>29.2</b>	<b>15.2</b>	17.6
○ Beam [2, 9, 24]	31.1	58.4	22.8	40.4	28.8	14.8	21.2
● Beam+INTER	<b>31.7</b>	54.8	<b>30.4</b>	<b>44.0</b>	26.0	<b>18.4</b>	16.4
○ VCD* [6]	30.3	54.4	24.4	38.0	26.0	16.8	22.4
● VCD*+INTER	<b>31.1</b>	<b>55.6</b>	<b>26.8</b>	<b>40.0</b>	<b>28.8</b>	15.2	20.4
○ OPERA <sup>†</sup> [11]	31.4	59.2	23.6	40.8	28.8	14.8	21.2
● OPERA <sup>†</sup> +INTER	<b>32.9</b>	56.8	<b>30.0</b>	<b>42.4</b>	<b>28.8</b>	<b>18.8</b>	20.8

Table 7. Validation of INTER on MMStar [3] using LLaVA-v1.5 (7B) [19].

Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. 2023. 2, 4, 8, 9, 10

- [6] Leng et al. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. *arXiv preprint arXiv:2311.16922*, 2023. 1, 2, 3, 4, 8, 10, 11, 12
- [7] Alessandro Favero, Luca Zancato, Matthew Trager, Siddharth Choudhary, Pramuditha Perera, Alessandro Achille, Ashwin Swaminathan, and Stefano Soatto. Multi-modal hallucination control by visual information grounding. In *Pro-*

method	Avg.	CP	FP	IR	LR	ST	MA
○ Nucleus [10]	30.5	50.4	24.8	40.0	27.6	17.6	22.4
● Nucleus+INTER	<b>32.3</b>	<b>53.2</b>	<b>25.2</b>	<b>42.0</b>	<b>31.6</b>	<b>18.0</b>	<b>23.6</b>
○ Greedy [23]	30.1	53.2	26.0	40.4	28.8	12.0	20.0
● Greedy+INTER	<b>33.3</b>	<b>55.2</b>	<b>29.6</b>	<b>42.4</b>	<b>30.8</b>	<b>17.2</b>	<b>24.8</b>
○ Beam [2, 9, 24]	30.7	52.4	26.0	42.0	28.8	12.0	23.2
● Beam+INTER	30.5	<b>52.8</b>	24.4	39.2	<b>29.6</b>	<b>14.0</b>	22.8
○ VCD* [6]	31.5	52.8	26.8	36.0	25.6	18.8	28.8
● VCD*+INTER	<b>32.3</b>	<b>52.8</b>	25.6	<b>39.6</b>	<b>32.0</b>	17.2	26.8
○ OPERA <sup>†</sup> [11]	30.5	52.8	26.4	41.6	28.4	12.0	22.0
● OPERA <sup>†</sup> +INTER	<b>30.9</b>	<b>52.8</b>	24.4	39.6	<b>29.6</b>	<b>14.8</b>	<b>24.0</b>

Table 8. Validation of INTER on MMStar [3] using mPLUG-owl2 [29].

model	method	CHAIR (512)		MMStar <sup>†</sup>	MME <sup>†</sup>
		$C_s \downarrow$	$C_I \downarrow$		
InstructBLIP [5]	M3ID [7]	63.1	21.1	29.8	1440.6
	Ritual [26]	62.1	20.9	29.5	1576.7
	SID [13]	59.7	21.4	28.1	1385.1
	<b>INTER (ours)</b>	<b>59.0</b>	<b>20.8</b>	<b>30.5</b>	<b>1595.5</b>
LLaVA-v1.5 [19]	M3ID [7]	67.1	19.7	30.9	1322.9
	Ritual [26]	52.4	15.8	31.2	1754.7
	SID [13]	52.0	14.3	31.0	1692.4
	<b>INTER (ours)</b>	<b>51.8</b>	<b>14.1</b>	<b>31.9</b>	1731.6

Table 9. Comparison with other methods.

ceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pages 14303–14312, 2024. 3, 8

- [8] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiaowu Zheng, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023. 1, 2, 3, 4, 10
- [9] Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012. 1, 2, 4, 8, 10, 11, 12
- [10] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2019. 2, 4, 8, 10, 11, 12
- [11] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multimodal 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, 2024. 1, 2, 3, 4, 8, 10, 11, 12
- [12] Drew A Hudson and Christopher D Manning. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709, 2019. 1, 2
- [13] Fushuo Huo, Wenchao Xu, Zhong Zhang, Haozhao Wang, Zhicheng Chen, and Peilin Zhao. Self-introspective decoding: Alleviating hallucinations for large vision-language models. *arXiv preprint arXiv:2408.02032*, 2024. 3, 8



method	InstructBLIP [5]		LLaVA-v1.5 (7B) [19]		Qwen-VL [1]		mPLUG-owl2 [29]	
	COCO [17]	AOKVQA [22]	COCO [17]	AOKVQA [22]	COCO [17]	AOKVQA [22]	COCO [17]	AOKVQA [22]
○ Greedy [23]	84.4	81.3	84.5	84.3	83.4	85.0	83.4	81.3
● Greedy+INTER	<b>85.1</b>	<b>83.7</b>	<b>86.4</b>	<b>84.4</b>	<b>86.0</b>	<b>86.4</b>	<b>83.4</b>	<b>81.7</b>

Table 10. **Evaluating the performance of INTER’s correction on Greedy Search [23]** by the mean F1-score across various partitions of two datasets in POPE [16]. Higher values are better.

Max Token	method	InstructBLIP [5]		LLaVA-v1.5 (7B) [19]		mPLUG-owl2 [29]	
		$C_S$	$C_I$	$C_S$	$C_I$	$C_S$	$C_I$
64	○ Greedy [23]	26.2	13.8	22.0	6.7	23.0	8.3
	● Greedy+INTER	<b>25.8</b>	<b>9.2</b>	<b>22.0</b>	<b>6.7</b>	<b>20.6</b>	<b>7.9</b>
512	○ Greedy [23]	49.2	21.9	48.8	13.4	58.2	18.5
	● Greedy+INTER	55.8	<b>18.1</b>	<b>48.8</b>	<b>13.4</b>	<b>54.4</b>	<b>17.9</b>

Table 11. **Evaluating the effectiveness of INTER in correcting Greedy Search** using LLaVA-v1.5 on CHAIR [21], with a maximum token length of 64 and 512. A smaller value indicates a lower degree of hallucinations.

model	RefCOCO [30]	
	testA	testB
DeepSeek-VL2-Tiny [27]	87.8	78.4
+INTER	<b>88.6</b>	<b>78.6</b>

Table 12. Performance on other types of tasks or different LVLMs.

- [14] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 3, 4
- [15] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 1
- [16] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023. 2, 9
- [17] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 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, 2014. 1, 2, 9
- [18] Haotian Liu. Llava-bench in the wild dataset. <https://huggingface.co/datasets/liuhaotian/llava-bench-in-the-wild>. Accessed: 2025-02-20. 2, 4
- [19] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2, 3, 4, 8, 9, 10, 11, 12
- [20] Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. Mmbench: Is your multi-modal model an all-around player? *arXiv preprint arXiv:2307.06281*, 2023. 1, 2, 8, 10, 11
- [21] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4035–4045, 2018. 1, 2, 9, 10
- [22] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. A-okvqa: A benchmark for visual question answering using world knowledge. In *European Conference on Computer Vision*, pages 146–162. Springer, 2022. 1, 2, 9
- [23] Yifan Song, Guoyin Wang, Sujian Li, and Bill Yuchen Lin. The good, the bad, and the greedy: Evaluation of llms should not ignore non-determinism. *arXiv preprint arXiv:2407.10457*, 2024. 2, 4, 8, 9, 10, 11, 12
- [24] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. 1, 2, 4, 8, 10, 11, 12
- [25] Xin Wang, Jie Ren, Shuyun Lin, Xiangming Zhu, Yisen Wang, and Quanshi Zhang. A unified approach to interpreting and boosting adversarial transferability. In *International Conference on Learning Representations*, 2020. 3
- [26] Sangmin Woo, Jaehyuk Jang, Donguk Kim, Yubin Choi, and Changick Kim. Ritual: Random image transformations as a universal anti-hallucination lever in lvlms. *arXiv preprint arXiv:2405.17821*, 2024. 3, 8
- [27] Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024. 3, 9
- [28] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, Chenliang Li, Yuanhong Xu, Hehong Chen, Junfeng Tian, Qian Qi, Ji Zhang, and Fei Huang. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*, 2023. 10
- [29] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13040–13051, 2024. 2, 4, 8, 9, 11, 12
- [30] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11–14,*

Datasets	Order	InstructBLIP [5]		LLaVA-v1.5 [19]		Qwen-VL [1]		mPLUG-owl2 [28]	
		Mean Absolute Value	Range	Mean Absolute Value	Range	Mean Absolute Value	Range	Mean Absolute Value	Range
MME [8]	$I(A \{v, p\})^{y_t}$	0.80	[-9.8, 5.7]	0.59	[-13.8, 5.3]	2.10	[-16.1, 10.0]	0.07	[-20.4, 27.8]
	$I(A \{p, v\})^{y_t}$	4.07	[-7.4, 22.8]	0.59	[-13.8, 9.3]	1.93	[-10.6, 8.2]	0.07	[-16.5, 10.2]
CHAIR [21]	$I(A \{v, p\})^{y_t}$	0.60	[-16.5, 14.2]	0.06	[-3.3, 3.9]	0.56	[-17.8, 12.4]	0.16	[-8.0, 7.7]
	$I(A \{p, v\})^{y_t}$	0.60	[-3.3, 3.9]	0.06	[-3.3, 3.9]	0.94	[-25.4, 8.1]	0.16	[-7.9, 7.5]

Table 13. The range of the metric  $I(A)^{y_t}$ .

method	Coarse Perception (CP)					Cross-instance Fine-grained Perception (FP-C)			Single-instance Fine-grained Perception (FP-S)			
	Image Emotion	Image Topic	Image Scene	Image Style	Image Quality	Action Recognition	Attribute Comparison	Spatial Relationship	Celebrity Recognition	Object Localization	Attribute Recognition	OCR
○ Nucleus [10]	71.0	76.4	94.1	66.5	24.7	82.3	44.0	11.9	75.5	25.7	73.5	53.2
● Nucleus+INTER	<b>77.5</b>	<b>80.7</b>	<b>96.1</b>	<b>70.8</b>	<b>25.3</b>	<b>84.7</b>	<b>51.1</b>	<b>13.0</b>	<b>80.1</b>	<b>35.2</b>	<b>81.4</b>	<b>59.0</b>
○ Greedy [23]	78.0	81.4	96.1	75.5	28.0	87.0	53.2	18.6	81.8	44.8	86.0	59.0
● Greedy+INTER	<b>78.0</b>	<b>81.4</b>	<b>96.1</b>	<b>75.5</b>	<b>28.0</b>	<b>87.0</b>	<b>53.2</b>	<b>18.6</b>	<b>81.8</b>	<b>44.8</b>	<b>86.0</b>	<b>59.0</b>
○ Beam [2, 9, 24]	78.0	81.4	96.1	75.5	28.0	87.0	53.2	18.6	81.8	44.8	86.0	59.0
● Beam+INTER	<b>78.0</b>	<b>81.4</b>	<b>96.1</b>	<b>75.5</b>	<b>28.0</b>	<b>87.0</b>	<b>53.2</b>	<b>18.6</b>	<b>81.8</b>	<b>44.8</b>	<b>86.0</b>	<b>59.0</b>
○ VCD* [6]	77.0	81.4	96.1	77.0	25.3	86.1	49.0	17.5	78.3	37.1	81.8	58.3
● VCD*+INTER	<b>77.0</b>	<b>82.1</b>	<b>96.1</b>	75.0	<b>28.0</b>	84.7	48.9	15.8	<b>80.3</b>	<b>37.8</b>	81.1	<b>60.9</b>
○ OPERA† [11]	78.0	81.4	96.1	75.5	28.0	87.0	53.2	18.6	81.8	43.5	86.0	59.0
● OPERA†+INTER	<b>78.0</b>	<b>81.4</b>	<b>96.1</b>	<b>75.9</b>	<b>28.0</b>	<b>87.0</b>	<b>53.2</b>	17.9	<b>81.8</b>	<b>43.5</b>	<b>86.0</b>	<b>59.0</b>

Table 14. Evaluating the performance of INTER on MM-Bench [20] using LLaVA-v1.5 (7B) [19], focusing on coarse perception and fine-grained perception subtasks.

method	Attribute Reasoning (AR)				Logic Reasoning (LR)			Relation Reasoning (RR)	
	Physical Property	Function Reasoning	Nature Relation	Future Prediction	Structuralized -Text	Image Understanding	Identity Reasoning	Social Relation	Physical Relation
○ Nucleus [10]	39.3	68.8	31.3	31.5	20.6	93.2	72.1	17.0	
● Nucleus+INTER	<b>44.3</b>	<b>77.3</b>	<b>38.0</b>	<b>39.2</b>	<b>24.1</b>	<b>93.8</b>	<b>85.5</b>	<b>22.3</b>	
○ Greedy [23]	43.8	82.9	34.6	41.5	27.0	95.5	86.1	25.5	
● Greedy+INTER	<b>43.8</b>	<b>82.9</b>	<b>34.6</b>	<b>41.5</b>	<b>27.0</b>	<b>95.5</b>	<b>86.1</b>	<b>25.5</b>	
○ Beam [2, 9, 24]	43.8	81.6	34.6	41.5	27.0	95.5	86.1	25.5	
● Beam+INTER	<b>43.8</b>	<b>82.2</b>	<b>34.6</b>	<b>41.5</b>	<b>27.0</b>	<b>95.5</b>	<b>86.1</b>	<b>25.5</b>	
○ VCD* [6]	41.1	73.7	39.1	40.0	23.1	95.5	83.7	18.1	
● VCD*+INTER	<b>41.1</b>	<b>73.7</b>	<b>39.1</b>	39.2	<b>23.4</b>	92.6	<b>88.4</b>	<b>22.3</b>	
○ OPERA† [11]	43.8	81.6	34.6	41.5	27.0	95.5	86.1	25.5	
● OPERA†+INTER	<b>43.8</b>	<b>81.6</b>	<b>34.6</b>	<b>41.5</b>	<b>27.0</b>	<b>95.5</b>	<b>86.1</b>	<b>25.5</b>	

Table 15. Evaluating the performance of INTER on MM-Bench [20] using LLaVA-v1.5 (7B) [19], focusing on attribute reasoning, logic reasoning and relation reasoning subtasks.

2016, *Proceedings, Part II 14*, pages 69–85. Springer, 2016.

9

- [31] Hao Zhang, Yichen Xie, Longjie Zheng, Die Zhang, and Quanshi Zhang. Interpreting multivariate shapley interactions in dnns. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10877–10886, 2021. 3

method	Coarse Perception (CP)					Cross-instance Fine-grained Perception (FP-C)			Single-instance Fine-grained Perception (FP-S)			
	Image Emotion	Image Topic	Image Scene	Image Style	Image Quality	Action Recognition	Attribute Comparison	Spatial Relationship	Celebrity Recognition	Object Localization	Attribute Recognition	OCR
○ Nucleus [10]	70.5	75.7	94.1	67.5	18.7	70.7	27.7	17.0	79.8	24.4	70.5	62.8
● Nucleus+INTER	<b>71.0</b>	<b>77.0</b>	<b>96.1</b>	<b>73.1</b>	<b>19.3</b>	<b>71.2</b>	<b>29.1</b>	<b>15.9</b>	<b>81.1</b>	<b>29.2</b>	<b>73.9</b>	<b>65.4</b>
○ Greedy [23]	<b>76.0</b>	<b>76.4</b>	<b>97.1</b>	<b>81.1</b>	<b>22.7</b>	<b>77.7</b>	<b>46.8</b>	<b>25.4</b>	<b>82.8</b>	<b>34.6</b>	<b>73.9</b>	<b>71.8</b>
● Greedy+INTER	<b>76.0</b>	73.6	96.1	<b>83.0</b>	12.0	<b>79.5</b>	41.8	18.6	81.8	28.6	<b>80.3</b>	<b>71.8</b>
○ Beam [2, 9, 24]	76.0	76.4	97.1	81.1	22.7	77.7	46.8	25.4	82.8	34.6	73.9	71.8
● Beam+INTER	<b>76.0</b>	<b>76.4</b>	<b>97.1</b>	<b>81.1</b>	<b>22.7</b>	<b>77.7</b>	<b>47.1</b>	<b>26.0</b>	<b>82.8</b>	<b>34.7</b>	<b>74.1</b>	<b>71.8</b>
○ VCD* [6]	71.5	79.3	96.3	75.9	17.3	72.6	32.6	17.5	79.8	26.7	75.0	66.7
● VCD*+INTER	<b>72.5</b>	<b>73.6</b>	<b>96.0</b>	<b>78.8</b>	<b>10.7</b>	<b>74.9</b>	<b>33.3</b>	<b>15.8</b>	<b>81.3</b>	<b>25.2</b>	<b>75.4</b>	<b>68.0</b>
○ OPERA <sup>†</sup> [11]	76.0	76.4	97.1	81.1	22.7	77.7	46.8	25.4	82.8	34.6	73.9	71.8
● OPERA <sup>†</sup> +INTER	<b>76.0</b>	<b>76.4</b>	97.0	<b>81.1</b>	<b>22.7</b>	<b>78.0</b>	<b>46.8</b>	<b>25.4</b>	82.7	<b>34.6</b>	<b>73.9</b>	<b>71.8</b>

Table 16. Evaluating the performance of INTER on MM-Bench [20] using mPLUG-owl2 [29], focusing on coarse perception and fine-grained perception subtasks.

method	Attribute Reasoning (AR)				Logic Reasoning (LR)		Relation Reasoning (RR)	
	Physical Property	Function Reasoning	Nature Relation	Future Prediction	Structuralized Image -Text Understanding	Identity Reasoning	Social Relation	Physical Relation
○ Nucleus [10]	32.4	75.3	33.0	36.2	18.8	92.1	73.3	22.3
● Nucleus+INTER	<b>33.3</b>	<b>76.6</b>	<b>36.3</b>	<b>40.0</b>	<b>20.9</b>	<b>93.8</b>	<b>79.1</b>	<b>26.9</b>
○ Greedy [23]	33.3	82.2	36.9	45.4	23.1	93.8	88.4	38.3
● Greedy+INTER	<b>36.1</b>	79.1	<b>41.3</b>	<b>46.9</b>	21.3	<b>96.0</b>	86.1	25.5
○ Beam [2, 9, 24]	33.3	82.2	36.9	45.4	24.5	93.8	88.4	34.0
● Beam+INTER	<b>33.9</b>	<b>82.2</b>	<b>36.9</b>	<b>45.4</b>	<b>24.5</b>	<b>93.8</b>	<b>88.4</b>	<b>38.2</b>
○ VCD* [6]	32.0	75.3	37.4	36.9	20.9	92.6	77.3	25.5
● VCD*+INTER	<b>34.3</b>	<b>76.0</b>	<b>39.3</b>	<b>43.9</b>	<b>20.2</b>	<b>95.5</b>	<b>79.1</b>	<b>20.2</b>
○ OPERA <sup>†</sup> [11]	<b>33.3</b>	<b>82.2</b>	<b>36.9</b>	<b>45.4</b>	<b>23.1</b>	<b>93.8</b>	<b>88.4</b>	<b>34.0</b>
● OPERA <sup>†</sup> +INTER	<b>33.8</b>	82.1	<b>36.9</b>	<b>45.4</b>	<b>24.5</b>	<b>93.8</b>	<b>88.9</b>	<b>38.3</b>

Table 17. Evaluating the performance of INTER on MM-Bench [20] using mPLUG-owl2 [29], focusing on attribute reasoning, logic reasoning and relation reasoning subtasks.

method	Coarse Perception (CP)			Fine-grained Perception (FP)			Instance Reasoning (IR)		
	Image Scene & Topic	Image Style & Quality	Image Emotion	Object Counting	Recognition	Localization	Single-Instance Reasoning	Cross-Instance Attribute Reasoning	Cross-Instance Relation Reasoning
○ Nucleus [10]	45.4	66.7	51.6	17.4	25.4	25.0	56.6	29.2	21.0
● Nucleus+INTER	<b>48.2</b>	<b>73.1</b>	<b>58.1</b>	<b>25.0</b>	<b>37.3</b>	<b>27.5</b>	51.5	<b>33.7</b>	<b>32.3</b>
○ Greedy [23]	48.9	74.4	67.7	20.7	29.7	20.0	52.5	31.5	32.3
● Greedy+INTER	48.2	66.7	58.1	<b>32.6</b>	27.1	<b>27.5</b>	<b>56.6</b>	<b>34.8</b>	<b>41.9</b>
○ Beam [2, 9, 24]	48.9	71.8	67.7	21.7	25.4	17.5	51.5	28.1	40.3
● Beam+INTER	47.5	69.2	<b>67.7</b>	<b>26.1</b>	<b>27.1</b>	<b>27.5</b>	<b>56.6</b>	<b>28.1</b>	<b>41.9</b>
○ VCD* [6]	44.0	70.5	61.3	27.2	22.9	22.5	48.5	32.6	29.0
● VCD*+INTER	<b>47.5</b>	<b>70.5</b>	54.8	26.1	<b>28.8</b>	<b>22.5</b>	<b>52.5</b>	31.5	<b>32.3</b>
○ OPERA <sup>†</sup> [11]	49.6	73.1	67.7	22.8	26.3	17.5	52.5	31.5	35.5
● OPERA <sup>†</sup> +INTER	45.4	71.8	<b>67.7</b>	<b>31.5</b>	<b>31.4</b>	<b>20.0</b>	<b>56.6</b>	<b>31.5</b>	<b>37.1</b>

Table 18. Evaluating the performance of INTER on MMStar [3] using LLaVA-v1.5 (7B) [19], focusing on coarse perception, fine-grained perception and instance reasoning subtasks.

method	Logit Reasoning (LR)			Science and Technology (ST)				Math (MA)		
	Code Sequence Reasoning	Diagram Reasoning	Common Reasoning	Biology & Chemistry & Physics	Electronics & Energy & Mechanical eng.	Geography & Earth Science & Agriculture	Geometry	Numeric & Calculation	Commonsense	Statistical Reasoning
○ Nucleus [10]	23.1	19.1	26.7	16.7	20.5	20.7	19.8	33.3		27.7
● Nucleus+INTER	<b>25.6</b>	<b>19.1</b>	<b>34.7</b>	<b>16.0</b>	17.9	<b>20.7</b>	<b>30.2</b>	31.3		19.3
○ Greedy [23]	23.1	22.7	33.7	12.5	15.4	17.2	25.6	27.1		16.9
● Greedy+INTER	<b>23.1</b>	21.8	<b>39.6</b>	<b>15.3</b>	10.3	<b>20.7</b>	17.4	18.8		<b>24.1</b>
○ Beam [2, 9, 24]	33.3	24.5	31.7	13.2	12.8	22.4	25.6	31.3		19.3
● Beam+INTER	<b>35.9</b>	21.8	<b>32.7</b>	<b>16.0</b>	<b>25.6</b>	<b>27.6</b>	24.4	27.1		<b>22.9</b>
○ VCD* [6]	20.5	20.0	34.7	18.1	15.4	17.2	27.9	20.8		26.5
● VCD*+INTER	<b>23.1</b>	<b>23.6</b>	<b>36.6</b>	17.4	7.7	<b>17.2</b>	17.4	<b>27.1</b>		<b>27.7</b>
○ OPERA <sup>†</sup> [11]	33.3	24.5	31.7	13.2	12.8	22.4	25.6	31.3		19.3
● OPERA <sup>†</sup> +INTER	<b>35.9</b>	21.8	<b>31.7</b>	<b>15.3</b>	<b>25.6</b>	<b>27.6</b>	23.3	27.1		<b>22.9</b>

Table 19. Evaluating the performance of INTER on MMStar [3] using LLaVA-v1.5 (7B) [19], focusing on logit reasoning, science technology and math capability subtasks.

method	Coarse Perception (CP)			Fine-grained Perception (FP)			Instance Reasoning (IR)		
	Image Scene & Topic	Image Style & Quality	Image Emotion	Object Counting	Recognition	Localization	Single-Instance Reasoning	Cross-Instance Attribute Reasoning	Cross-Instance Relation Reasoning
○ Nucleus [10]	43.9	56.4	64.5	27.2	24.6	20.0	49.5	28.1	41.9
● Nucleus+INTER	<b>47.5</b>	<b>58.9</b>	<b>70.9</b>	<b>31.5</b>	<b>27.9</b>	15.0	<b>53.5</b>	<b>31.5</b>	30.6
○ Greedy [23]	46.1	59.0	71.0	26.1	30.5	12.5	53.5	30.3	33.9
● Greedy+INTER	<b>48.9</b>	<b>61.5</b>	67.7	<b>33.7</b>	29.7	<b>20.0</b>	<b>55.6</b>	<b>32.6</b>	<b>35.5</b>
○ Beam [2, 9, 24]	45.4	57.7	70.9	27.2	29.7	12.5	53.5	29.2	41.9
● Beam+INTER	<b>46.1</b>	<b>57.7</b>	<b>70.9</b>	<b>28.2</b>	26.3	10.0	<b>53.5</b>	28.1	32.3
○ VCD* [6]	46.8	58.9	64.5	27.2	28.2	20.0	46.5	28.1	30.6
● VCD*+INTER	<b>47.5</b>	56.4	<b>67.7</b>	<b>28.3</b>	24.6	<b>22.5</b>	<b>51.5</b>	<b>30.3</b>	<b>33.9</b>
○ OPERA <sup>†</sup> [11]	45.4	58.9	70.9	27.2	29.7	15.0	53.5	29.2	40.3
● OPERA <sup>†</sup> +INTER	<b>46.1</b>	57.7	70.9	<b>28.3</b>	26.3	10.0	<b>53.5</b>	28.1	33.9

Table 20. Evaluating the performance of INTER on MMStar [3] using mPLUG-owl2 [29], focusing on coarse perception, fine-grained perception and instance reasoning subtasks.

method	Logit Reasoning (LR)			Science and Technology (ST)				Math (MA)		
	Code Sequence Reasoning	Diagram Reasoning	Common Reasoning	Biology & Chemistry & Physics	Electronics & Energy & Mechanical eng.	Geography & Earth Science & Agriculture	Geometry	Numeric & Calculation	Commonsense	Statistical Reasoning
○ Nucleus [10]	25.6	21.8	34.7	15.1	21.7	20.7	22.4	22.9		22.1
● Nucleus+INTER	12.8	18.2	<b>42.6</b>	<b>19.2</b>	<b>36.9</b>	<b>24.1</b>	<b>19.8</b>	<b>20.8</b>		<b>26.7</b>
○ Greedy [23]	30.8	24.6	32.7	11.0	8.7	17.2	18.1	25.0		19.8
● Greedy+INTER	20.5	<b>24.6</b>	<b>41.6</b>	<b>15.8</b>	<b>23.9</b>	15.5	<b>19.0</b>	<b>27.1</b>		<b>31.4</b>
○ Beam [2, 9, 24]	25.6	25.5	33.7	11.6	6.5	17.2	18.9	25.0		27.9
● Beam+INTER	<b>25.6</b>	24.5	<b>36.6</b>	<b>12.3</b>	<b>15.2</b>	<b>17.2</b>	<b>19.8</b>	22.9		26.7
○ VCD* [6]	28.2	20.9	29.7	23.3	10.9	13.8	23.3	33.7		33.7
● VCD*+INTER	17.9	<b>25.5</b>	<b>44.6</b>	15.8	<b>21.7</b>	<b>17.2</b>	<b>25.0</b>	27.1		29.1
○ OPERA <sup>†</sup> [11]	25.6	23.6	34.7	11.6	6.5	17.2	19.0	25.0		24.4
● OPERA <sup>†</sup> +INTER	<b>25.6</b>	<b>25.5</b>	<b>35.6</b>	<b>13.0</b>	<b>17.4</b>	<b>17.2</b>	<b>21.6</b>	<b>25.0</b>		<b>26.7</b>

Table 21. Evaluating the performance of INTER on MMStar [3] using mPLUG-owl2 [29], focusing on logit reasoning, science, technology, and math capability subtasks.