Mitigating Object Hallucinations in Large Vision-Language Models with Assembly of Global and Local Attention

Supplementary Material

A1. Limitation and Future Work

Despite the superb performance in mitigating hallucinations and enhancing the general perception capabilities of LVLMs, our work could be improved in several aspects. First, we conducted experiments on the most widely used LVLMs due to resource constraints. It will be useful to evaluate our model on larger LVLMs such as LLaVA 34B and Flamingo 70B [4]. In addition, this work focuses on text and image data. It could be extended to data from other modalities such as videos. We will examine these problems in our future work.

A2. Evaluation Metrics

The details of evaluation metrics for different datasets are listed below.

POPE. Since the output of the model for the POPE dataset [40] is limited to two types (Yes or No), it is convenient to measure the model performance with binary classification metrics: Accuracy, Precision, Recall, and F1 score.

ROPE. We utilize the official code of ROPE [10] to measure the multi-object hallucination mitigation performance of different models. Similar to POPE, the evaluation metrics include Accuracy, Precision, Recall, and F1 score.

MME. Similar to the POPE dataset, the MME dataset [19] contains only two types of answers (*i.e.*, Yes or No). Following the setting in their original paper, we use the sum of accuracy and accuracy+ as the final score, where accuracy is calculated based on each question and accuracy+ is calculated based on each image where both of the two questions need to be answered correctly. So accuracy+ is a stricter measurement that can better reflect the comprehensive understanding degree of the model.

CHAIR. Different from discriminative datasets like POPE and MME, CHAIR [52] (*i.e.*, Caption Hallucination Assessment with Image Relevance) is a framework that quantifies object hallucinations for generative datasets. The sentence-level score CHAIR$_S$ represents the proportion of generated captions that contain hallucinations, while the instance-level score CHAIR$_I$ denotes the proportion of hallucinated objects relative to all mentioned objects in the generated captions. In addition, we also evaluate the semantic detailedness of generated captions with the metric Recall. CHAIR$_S$, CHAIR$_I$, and Recall are computed as follows.

$$C_S = \frac{|\{\text{Captions with hallucinated objects}\}|}{|\{\text{All captions}\}|} \tag{A1}$$

$$C_I = \frac{|\{\text{Hallucinated objects}\}|}{|\{\text{All mentioned objects}\}|} \tag{A2}$$

$$Recall = \frac{|\{\text{Accurate objects}\}|}{|\{\text{Ground-truth objects}\}|} \tag{A3}$$

LLaVA-Bench-Wild. LLaVA-Bench-Wild [45] contains 24 images with 60 questions to assess the capability of LVLMs in tackling challenging tasks and their adaptability to new domains. Following previous works [26, 33], we use GPT-4 to evaluate the accuracy and detailedness of the generated captions. Specifically, the metric Accuracy measures the captions' alignment with the image content and the metric Detailedness gauges the richness of details in the generated captions. The specific configurations used to prompt GPT-4 are listed in Table A10.

A3. Experimental Settings

For the POPE [40] and ROPE [10] datasets, α is set to 2 and β is set to 0.5. For the MME [19] dataset, we set α to 2 and β to 0.5 for LLaVA-1.5, while setting α and β to 0.1 for InstructBLIP. For the CHAIR [52] and LLaVA-Bench-Wild [45] datasets, α is set to 2 and β is set to 0.5. All 7B model experiments are conducted on a single RTX 3090 GPU, and all 13B model experiments are conducted on two RTX 3090 GPUs.

A4. Different Masking Strategies

In our paper, we mask the image at the pixel level by setting the RGB value of the masking pixel to 0. "Patch-level" masking means masking the image patches based on the activation value of the patch regions. "Soft" masking involves multiplying the GradCAM activation value by the original pixel values instead of setting the masking pixels to 0. "Feature-level" masking means masking the image features extracted by the vision encoder rather than masking the input image. "Random" masking means randomly masking the pixels of the input images. The results of different masking strategies are listed in Table 5.

A5. Detailed Results on POPE

The results of LLaVA-1.5 7B [45] and InsturctBLIP 7B [15] on the full POPE dataset [40] with three subsets (*i.e.*, MSCOCO [42], A-OKVQA [53] and GQA [27]) and three negative sample settings (*i.e.*, Random, Popular and Adversarial) are listed in Table A1. From the table, we can see that the proposed decoding strategy *AGLA* consistently outperforms the regular decoding strategy by large margins

Table A1. Results on the full POPE dataset with LLaVA-1.5 7B [45] and InstructBLIP 7B [15].

Dataset	**Setting**	**Model**	**Decoding**	Accuracy ↑	Precision	Recall	F1 Score ↑
COCO	*Random*	LLaVA-1.5 [45]	Regular	83.29	92.13	72.80	81.33
			AGLA	**87.46**	97.52	76.87	**85.97**
		InstructBLIP [15]	Regular	80.71	81.67	79.19	80.41
			AGLA	**87.63**	93.88	80.51	**86.68**
	Popular	LLaVA-1.5 [45]	Regular	81.88	88.93	72.80	80.06
			AGLA	**86.12**	94.33	76.87	**84.71**
		InstructBLIP [15]	Regular	78.22	77.87	78.85	78.36
			AGLA	**84.63**	87.75	80.51	**83.97**
	Adversarial	LLaVA-1.5 [45]	Regular	78.96	83.06	72.75	77.57
			AGLA	**83.87**	89.48	76.76	**82.63**
		InstructBLIP [15]	Regular	75.84	74.30	79.03	76.59
			AGLA	**81.90**	82.76	80.58	**81.66**
VQA	*Random*	LLaVA-1.5 [45]	Regular	83.45	87.24	78.36	82.56
			AGLA	**89.28**	93.18	84.76	**88.77**
		InstructBLIP [15]	Regular	80.91	77.97	86.16	81.86
			AGLA	**87.80**	86.76	89.22	**87.97**
	Popular	LLaVA-1.5 [45]	Regular	79.90	80.85	78.36	79.59
			AGLA	**85.63**	86.27	84.76	**85.51**
		InstructBLIP [15]	Regular	76.19	72.16	85.28	78.17
			AGLA	**82.27**	78.33	89.22	**83.42**
	Adversarial	LLaVA-1.5 [45]	Regular	74.04	72.08	78.49	75.15
			AGLA	**78.85**	75.83	84.71	**80.03**
		InstructBLIP [15]	Regular	70.71	65.91	85.83	75.56
			AGLA	**74.79**	69.33	88.91	**77.91**
GQA	*Random*	LLaVA-1.5 [45]	Regular	83.73	87.16	79.12	82.95
			AGLA	**88.89**	92.53	84.60	**88.39**
		InstructBLIP [15]	Regular	79.65	77.14	84.29	80.56
			AGLA	**86.46**	85.84	87.31	**86.57**
	Popular	LLaVA-1.5 [45]	Regular	78.17	77.64	79.12	78.37
			AGLA	**83.67**	83.05	84.60	**83.82**
		InstructBLIP [15]	Regular	73.87	69.63	84.69	76.42
			AGLA	**78.67**	74.44	87.31	**80.36**
	Adversarial	LLaVA-1.5 [45]	Regular	75.08	73.19	79.16	76.06
			AGLA	**80.66**	78.30	84.82	**81.43**
		InstructBLIP [15]	Regular	70.56	66.12	84.33	74.12
			AGLA	**75.18**	70.19	87.53	**77.91**

(average 5.5% accuracy and 5.1% F1 improvement) on all LVLMs. The improved performance on popular and adversarial (*i.e.*, co-occurrence) settings of different subsets validates that our model can better mitigate the statistical bias of LVLMs by addressing the *attention deficiency* issue.

A6. Results on POPE with More LVLMs

In addition to LLaVA-1.5 [45] and InsturctBLIP [15], we perform experiments with more kinds of LVLMs, such as Qwen-VL [7] and MiniCPM-V [71]. The results are shown in Table A2 and A3, respectively. From the tables, we can see that the proposed *AGLA* consistently outperforms the regular decoding strategy by large margins on all evaluation metrics and LVLMs, which can show the generalizability of our method towards different kinds of LVLMs.

A7. Results on POPE when LVLMs Scale Up

Table A4 and A5 show experimental results on the POPE dataset [40] when LLaVA-1.5 [45] and InstructBLIP [15] extend to larger 13B variants. Notably, *AGLA* consistently boosts model performance in all subsets and settings of the POPE dataset [40], confirming its robustness towards different model scales.

Table A2. Results on the full POPE dataset with QWen-VL [7].

Dataset	**Setting**	**Decoding**	Accuracy ↑	Precision	Recall	F1 Score ↑
COCO	*Random*	Regular	83.17	96.80	68.60	80.30
		AGLA	**84.60**	98.23	70.47	**82.07**
	Popular	Regular	82.76	95.63	68.66	79.94
		AGLA	**84.40**	97.16	70.87	**81.95**
	Adversarial	Regular	82.00	92.47	69.67	79.47
		AGLA	**82.70**	93.22	70.53	**80.30**
VQA	*Random*	Regular	84.53	93.31	74.40	82.79
		AGLA	**86.77**	95.02	77.60	**85.43**
	Popular	Regular	84.43	92.44	75.00	82.81
		AGLA	**86.30**	94.09	77.47	**84.97**
	Adversarial	Regular	78.20	80.61	74.27	77.31
		AGLA	**80.73**	82.79	77.60	**80.11**
GQA	*Random*	Regular	81.20	90.48	69.73	78.76
		AGLA	**83.90**	93.05	73.26	**81.98**
	Popular	Regular	78.23	82.90	71.13	76.56
		AGLA	**80.80**	85.70	73.93	**79.38**
	Adversarial	Regular	75.63	79.06	69.73	74.10
		AGLA	**78.73**	82.16	73.40	**77.53**

Table A3. Results on the full POPE dataset with MiniCPM-V [71].

Dataset	**Setting**	**Decoding**	Accuracy ↑	Precision	Recall	F1 Score ↑
COCO	*Random*	Regular	86.00	87.09	84.53	85.79
		AGLA	**88.90**	92.50	84.67	**88.41**
	Popular	Regular	84.00	83.20	85.20	84.19
		AGLA	**85.76**	86.62	84.60	**85.60**
	Adversarial	Regular	80.27	77.99	84.33	81.03
		AGLA	**82.26**	80.86	84.53	**82.66**
VQA	*Random*	Regular	84.70	81.56	89.66	85.43
		AGLA	**88.03**	85.36	91.80	**88.47**
	Popular	Regular	79.30	74.40	89.33	81.19
		AGLA	**82.30**	77.08	91.93	**83.85**
	Adversarial	Regular	72.33	66.43	90.26	76.54
		AGLA	**73.80**	67.53	91.66	**77.77**
GQA	*Random*	Regular	84.63	81.89	88.93	85.27
		AGLA	**87.06**	84.41	90.93	**87.54**
	Popular	Regular	74.93	69.33	89.40	78.10
		AGLA	**77.26**	71.52	90.60	**79.94**
	Adversarial	Regular	71.46	65.76	89.53	75.83
		AGLA	**72.90**	66.89	90.66	**76.99**

A8. Results on Perception-related MME

To validate the effectiveness of our model for enhancing general perception capability for LVLMs, we perform experiments on the perception-related tasks of the MME benchmark [19]. As shown in Table A6, our model obtains much better performance than regular baselines and achieves uniformly improvement than previous state-of-the-art models. The improvement may come from the ensemble of local and global image attention to capture both discrim-

Table A4. Results on the full POPE dataset with LLaVA-1.5 13B [45].

Dataset	Setting	Decoding	Accuracy ↑	Precision	Recall	F1 Score ↑
COCO	*Random*	Regular	83.31	91.46	73.48	81.49
		AGLA	**87.26**	97.94	76.13	**85.67**
	Popular	Regular	82.47	89.55	73.53	80.75
		AGLA	**86.46**	96.12	76.00	**84.88**
	Adversarial	Regular	80.00	84.46	73.53	78.62
		AGLA	**84.56**	91.78	75.93	**83.11**
VQA	*Random*	Regular	83.60	86.95	79.07	82.82
		AGLA	**89.40**	94.57	83.60	**88.74**
	Popular	Regular	81.16	82.53	79.06	80.76
		AGLA	**86.93**	89.57	83.60	**86.48**
	Adversarial	Regular	76.43	74.79	79.73	77.18
		AGLA	**80.86**	79.22	83.67	**81.38**
GQA	*Random*	Regular	84.50	87.52	80.46	83.84
		AGLA	**89.43**	93.84	84.40	**88.87**
	Popular	Regular	80.67	80.78	80.46	80.62
		AGLA	**86.80**	88.65	84.40	**86.47**
	Adversarial	Regular	77.10	75.29	80.66	77.88
		AGLA	**82.53**	81.32	84.46	**82.86**

Table A5. Results on the full POPE dataset with InstructBLIP 13B [15].

Dataset	Setting	Decoding	Accuracy ↑	Precision	Recall	F1 Score ↑
COCO	*Random*	Regular	82.50	86.35	77.20	81.52
		AGLA	**88.67**	95.81	80.87	**87.71**
	Popular	Regular	79.30	81.64	75.60	78.50
		AGLA	**85.36**	88.86	80.86	**84.67**
	Adversarial	Regular	75.96	76.44	75.06	75.74
		AGLA	**82.67**	84.02	80.66	**82.31**
VQA	*Random*	Regular	82.16	82.05	82.33	82.19
		AGLA	**89.53**	90.06	88.86	**89.46**
	Popular	Regular	76.96	73.89	83.40	78.35
		AGLA	**83.73**	80.59	88.86	**84.52**
	Adversarial	Regular	72.23	67.74	84.86	75.34
		AGLA	**75.66**	70.11	89.46	**78.61**
GQA	*Random*	Regular	81.53	81.70	81.26	81.48
		AGLA	**87.60**	88.63	86.26	**87.43**
	Popular	Regular	71.86	68.82	79.93	73.96
		AGLA	**79.10**	75.45	86.26	**80.49**
	Adversarial	Regular	71.60	67.76	82.40	74.36
		AGLA	**74.43**	69.88	85.86	**77.05**

inative and generative features, which is important for enhancing the visual perceptual abilities of LVLMs.

A9. Effect of Hyper-parameters

Effect of Weighting Factor. We investigate the influence of the weighting factor α in Eq. 3 on the POPE-COCO dataset with the adversarial setting in Table A7. Increasing α is equivalent to amplifying the importance of discriminative local features from augmented images, which is important for mitigating object hallucinations, so our model gets consistent improvement when α increases. On the other hand, increasing α can be viewed as applying a smaller tempera-

Table A6. Results on MME perception-related tasks.

Model	Decoding	*Existence*	*Count*	*Position*	*Color*	*Posters*	*Celebrity*	*Scene*	*Landmark*	*Artwork*	*OCR*	***Perception Total***
LLaVA-1.5 [45]	Regular	175.67	124.67	114.00	151.00	127.82	113.59	148.30	129.95	102.20	92.00	1279.19
	DOLA	175.00	108.33	90.00	138.33	121.43	108.82	146.50	124.12	107.50	112.50	1232.54
	OPERA	180.67	133.33	123.33	155.00	134.69	116.76	152.75	133.01	103.25	100.00	1332.79
	VCD	184.66	138.33	128.67	153.00	132.11	120.94	152.20	140.45	109.60	104.00	1363.96
	AGLA	**195.00**	**153.89**	**129.44**	**161.67**	**137.07**	**126.96**	**157.42**	**160.13**	**116.08**	**135.00**	**1472.66**
InstructBLIP [15]	Regular	161.67	62.22	63.33	114.44	120.41	94.80	143.00	131.74	96.92	80.83	1069.36
	DOLA	175.00	50.00	46.67	113.33	130.27	96.18	148.25	131.41	99.25	72.50	1062.86
	OPERA	175.00	55.00	50.00	118.15	122.86	80.00	149.25	138.79	90.75	65.00	1044.80
	VCD	170.00	61.67	57.22	114.44	121.09	**104.41**	140.75	**140.96**	**103.08**	**82.50**	1096.12
	AGLA	**180.00**	**63.33**	**65.56**	**119.44**	**130.38**	96.57	**150.58**	135.76	97.50	70.00	**1109.12**

Table A7. An ablation study on different weighting factors α with LLaVA-1.5 7B [45].

α	Accuracy ↑	Precision	Recall	F1 Score ↑
0.5	81.80	86.30	75.60	80.60
1.0	82.83	87.80	76.27	81.63
1.5	83.40	88.66	76.60	82.19
2.0	83.80	89.12	77.00	82.62
3.0	84.11	88.53	78.37	83.14

Table A8. An ablation study on different constraint factors β with LLaVA-1.5 7B [45].

β	Accuracy ↑	Precision	Recall	F1 Score ↑
0.01	84.73	89.35	78.87	83.78
0.1	84.11	88.53	78.37	83.14
0.2	83.90	89.39	76.93	82.69
0.5	84.00	89.41	77.13	82.82
1.0	83.53	89.24	76.27	82.24

ture to sampling decoding, which can skew the distribution towards high-probability tokens and avoid selecting low-probability random tokens. Nevertheless, introducing augmented images for decoding with different α consistently outperforms the regular decoding, which can validate the stability of our model in mitigating object hallucinations.

Effect of Adaptive Plausible Constraint Factor. We investigate the influence of the adaptive plausible constraint factor β in Eq. 4 on the POPE-COCO dataset with the adversarial setting in Table A8. Larger β indicates more aggressive truncation, keeping only high-probability tokens. The table illustrates that our model is robust to the change of β. However, the factor is important for generative tasks to avoid generated tokens being dominated by local image features from the augmented images and compromising global perception ability from the original images.

A10. Qualitative Results for Different Queries

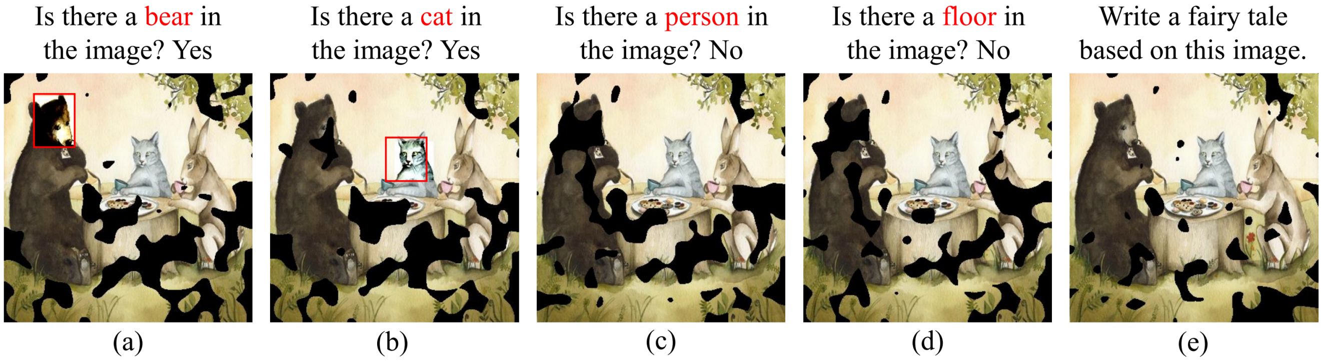
Fig. A1 shows the masked images generated by our Image-Prompt Matching (IPM) module with different prompt queries. First, for questions that result in "yes" answers (*i.e.*, Fig. (a) and (b)), the matching model tries to match the objects in the input queries and mask out other irrelevant backgrounds to avoid interference. Second, for questions that result in "no" answers (*i.e.*, Fig. (c) and (d)), the matching model fails to match any objects and tends to match background areas. So our model can mask some irrelevant objects to mitigate object hallucination caused by object co-occurrence association. Third, for generative tasks with general prompts that contain no objects (*i.e.*, Fig. (e)), our matching model can still identify areas of interest (*e.g.*, animals and the table) due to the pre-training task on caption generation and mask irrelevant backgrounds, which can mitigate interference and prevent loss of details in generated captions. In summary, IPM is effective for different kinds of input queries and helps to mitigate object hallucinations caused by association.

A11. Qualitative Results for Multiple-object Queries

In addition to experiments on the ROPE dataset [10], Fig. A2 also illustrates that GradCAM can work well when queries contain multiple objects, showing the effectiveness of our model towards multiple-object hallucination mitigation. Furthermore, the experiments also demonstrate that GradCAM can detect general regions of interest rather than a single object, which makes our model competent for a wide range of general tasks in the real world, validated by the improvement over the MME dataset in Table A6.

A12. Prompt for GPT-4 Aided Evaluation

To evaluate model performance on caption generation, we use GPT-4 to assess the accuracy and detailedness of LVLMs' responses, following previous works [26, 33]. The prompt used for GPT-4 is detailed in Table A10. Further-


Is there a bear in the image? Yes
Is there a cat in the image? Yes
Is there a person in the image? No
Is there a floor in the image? No
Write a fairy tale based on this image.

(a) (b) (c) (d) (e)

Figure A1. Masked images with different prompt queries. (a) and (b) are prompts that contain objects in the image, where the detected objects are marked in red. (c) and (d) are prompts that do not contain objects in the image. (e) is a general prompt.

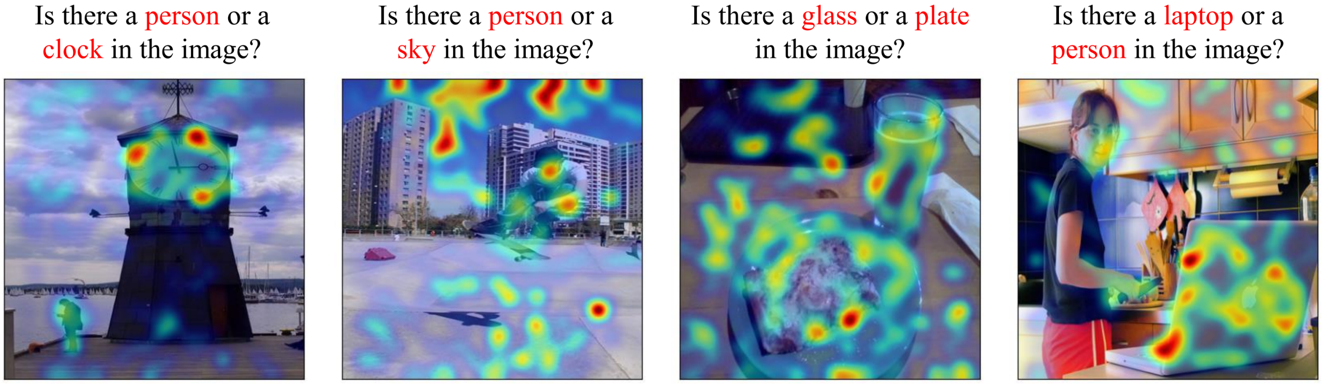

Is there a person or a clock in the image?
Is there a person or a sky in the image?
Is there a glass or a plate in the image?
Is there a laptop or a person in the image?

Figure A2. GradCAM results when queries contain multiple objects.

Table A9. Model performance and inference time on the POPE dataset with LLaVA-1.5 [45].

	VCD	OPERA	AGLA-small	AGLA
F1 Score	83.16	83.55	84.11	84.58
Infer. Time	0.56s	1.64s	0.63s	0.69s

more, we also present an illustrative example in Fig. A3 for a better understanding of the evaluation process of GPT-4.

A13. Efficiency

We benchmark different models with F1 scores and inference time per sample on the POPE dataset with LLaVA-1.5 [45] in Table A9. Experimental results show that our model performs the best with slight inference overhead. We also test *AGLA-small*, an *AGLA* variant using a much smaller matching model with 120M parameters, which is more efficient but achieves a competitive F1 score as well.

A14. Examples for Caption Generation

In order to demonstrate the quality of generated responses by different models more clearly, we present more qualitative results on the CHAIR evaluation [52] in Fig. A4 and A5. From the results, we can see that our model can generate captions with fewer object hallucinations, without loss of detailedness of the captions, which is consistent with the results in Table 3 and 4.

Table A10. The prompt for GPT-4 to evaluate captions with Accuracy and Detailedness, following previous works [26, 33].

Description:
AI that scores image description accuracy and detailedness.

Instructions:

You are an AI designed to evaluate and score the performance of two AI assistants in describing a given image. Your primary focus is on the accuracy and detailedness of their descriptions. You will assess the accuracy by checking for hallucinations - any part of the description that is inconsistent with the image content. For detailedness, you will consider how rich the response is in necessary details, excluding any hallucinated parts. You will provide scores on a scale from 1 to 10 for each assistant separately, based on these criteria. After scoring, you will offer an explanation for your evaluation, ensuring it is free from bias and not influenced by the order of presentation of the responses.

Input format:

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

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

Output format:

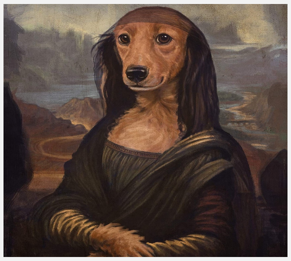
Accuracy:
Scores of the two answers:
Reason:

Detailedness:
Scores of the two answers:
Reason:

[Prompt]

Please describe this image in detail.

[Image]



[Assistant 1]

The image features a large dog sitting in front of a full-length, highly detailed portrait of itself. The dog wears a bandana and appears to be the main subject of the painting. It seems to be a captivating artwork that captures the unique and intricate beauty of the dog.\n\n In the background, a person can also be seen standing slightly to the left of the dog, possibly observing or admiring the painting. The scene creates an interesting contrast between the real-life dog and its depiction in the artwork.

[End of Assistant 1]

[Assistant 2]

The image features a dog painted to resemble a human woman. The dog is wearing a dress and a scarf, giving it the appearance of a Renaissance portrait. It is positioned in front of a landscape background that includes a water body and a mountain, adding an artistic touch to the overall scene. The dog's body language suggests that it is posing for the portrait, creating a captivating visual effect.

[Assistant 2]

[GPT-4 Output]

Accuracy:
Scores of the two answers:
Assistant 1: 1
Assistant 2: 9

Reason:
Assistant 1 inaccurately describes the image, hallucinating a live dog in front of a portrait and a person in the background, none of which exist in the image. This significant misrepresentation results in a very low accuracy score.
Assistant 2 accurately identifies that the image is of a dog painted in the style of a human Renaissance portrait, correctly noting the attire and setting, aligning well with the content of the image.

Detailedness:
Scores of the two answers:
Assistant 1: 2
Assistant 2: 8

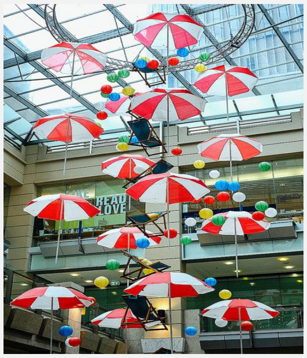
Reason:
Assistant 1 provides detail in describing elements that do not exist in the image, such as a live dog, a bandana, and a background person. The non-existent details are irrelevant to the actual image content, thereby reducing the score for detailedness despite the descriptive attempt.
Assistant 2 offers a more detailed description of the actual image, noting the dog's dress and scarf in a Renaissance style and the landscape background. The details about the dog's body language and the scene add to the richness of the description, enhancing its detailedness score.

[End of GPT-4 Output]

Figure A3. An example of the GPT-4 evaluation in open-ended generation tasks. “Assistant 1” is Regular decoding and “Assistant 2” is the proposed model *AGLA*.

[Prompt]
Please describe this image in detail.

[Image]



[Regular]
The image features a unique, colorful display of numerous umbrellas hanging from the ceiling of a building, likely a store or a public space. The umbrellas are in various shapes and sizes, creating a vibrant and eye-catching scene.\n\n Apart from the umbrellas, there are also several ***balls*** scattered throughout the display, adding more color and visual interest to the scene. The ***balls*** are of different shapes and sizes, further enhancing the overall aesthetic.\n\n The umbrellas are suspended at different heights, with some closer to the bottom and others higher up. This artistic arrangement gives the impression of a festive or whimsical atmosphere, attracting attention and making it a memorable experience for visitors.

[VCD]
The image presents a sky full of various luminous umbrellas suspended under a glass roof, creating a colorful atmosphere. There are different sizes and colors of umbrellas, adding to the vibrant, whimsical display. Beneath the umbrellas, the shopping floor exudes a luminous interior design.\n\n There are a few ***people*** scattered around the scene, possibly passersby or visitors to the shop beneath the roof. The brightly colored umbrellas and glass roof adorn the interior of the mall, giving it an artistic and captivating vibe.

[OPERA]
The image features a large indoor space filled with numerous colorful umbrellas suspended from the ceiling. These umbrellas come in various sizes and colors, creating a vibrant and eye-catching display. The umbrellas are spread throughout the room, with some closer to the foreground and others further in the background.\n\n In addition to the umbrellas, there are several chairs scattered around the room, some ***people*** sit and enjoy the unique atmosphere created by the umbrellas. The combination of the umbrellas and chairs creates a visually appealing and inviting environment.

[DOLA]
The image showcases a large and colorful installation of umbrellas suspended from the ceiling of a building, creating an artistic and vibrant atmosphere. The umbrellas are of various sizes and colors, adding a lively touch to the scene. In total, there are ***10 umbrellas*** visible in the image, with some placed higher and others hanging lower.\n\n A few chairs can also be seen scattered throughout the scene, possibly serving as a part of the installation or simply placed for decoration. These chairs are positioned at different heights and angles, adding more visual interest to the overall display.

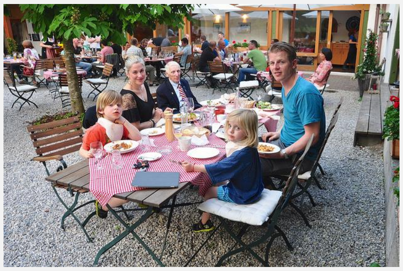
[AGLA]
The image features a large indoor space with a unique and colorful installation of numerous umbrellas hanging from the ceiling. The umbrellas are of various sizes and colors, creating an eye-catching display. Some umbrellas are red, and they are all suspended from the ceiling at different heights.\n\n In addition to the umbrellas, there are several chairs scattered throughout the space, adding to the overall aesthetic. The chairs are positioned at different angles and heights, contributing to the dynamic and visually interesting atmosphere of the room.

Figure A4. An example of generated captions by different decoding methods. Hallucinated content is marked in ***red***.

[Prompt]

Please describe this image in detail.

[Image]



[Regular]

The image showcases a family gathered around a dining table outside, enjoying a meal together. There are multiple chairs placed around the table, and several family members can be seen sitting on them. Various people can be seen sitting or standing in the background as well.\n\n The table is adorned with a mix of dishes and cutlery, such as wine glasses, cups, ***forks***, and ***knives***. Several ***bottles*** are also scattered around the table, suggesting different beverages being served.\n\n The scene is set in an outdoor area, possibly in a park or similar environment, adding to the family's enjoyment of their time together.

[VCD]

The scene features a crowded restaurant patio area lined with several dining tables, each with chairs arranged around them. Two poofy chairs are prominently visible placing a family, consisting of a father, two young children, and a grand lady, at one of the tables. They seem to be enjoying a meal with utensils like ***forks***, ***knives***, and cups laid out.\n\n In the vicinity, there's an array of other people eating, conversing or engaged in other daily activities. Some can be seen near other dining tables, while others are sitting or standing around the area. The restaurant itself has a few ***potted plants*** inside as well as a collection of wine glasses, giving an impression of a lively and pleasant atmosphere.

[OPERA]

The image depicts a group of people, including a man, a woman, and two children, sitting around a dining table outfitted with a red and white checkered tablecloth. They are enjoying a meal together, with plates of food and wine ***bottles*** placed on the table.\n\n There are several chairs surrounding the table, and a few more people can be seen in the background, likely engaged in conversation or waiting for their turn to join the meal. The dining area is furnished with additional chairs and dining tables, creating a cozy and inviting atmosphere for the family gathering.

[DOLA]

The image depicts a lively outdoor restaurant setting with several people dining together at picnic tables. A man, a woman, and two children are seated at a dining table, sharing a meal together. The table is set with various items such as ***forks***, ***spoons***, ***knives***, and wine glasses. There are also cups, bowls, and a couple of ***bottles*** placed on the table.\n\n The dining area is surrounded by numerous chairs, and there are additional tables and benches spread throughout the scene. The restaurant is bustling with activity, and people can be seen at different locations, either sitting at tables or walking around. Overall, it's a warm and inviting atmosphere, showcasing a family enjoying a meal in a relaxed and convivial setting.

[AGLA]

The image features a family sitting around a dining table in an outdoor restaurant. They are enjoying a meal together, with various plates of food, cups, and wine glasses spread across the table.\n\n There are multiple chairs surrounding the table, with some placed close to the family and others further away. A few other people can be seen in the background, likely other diners at the restaurant. The atmosphere appears to be casual and pleasant, as the family enjoys their time together.

Figure A5. An example of generated captions by different decoding methods. Hallucinated content is marked in ***red***.