OPERA: Alleviating Hallucination in Multi-Modal Large Language Models via Over-Trust Penalty and Retrospection-Allocation Supplementary Material

Qidong Huang[1,2,*], Xiaoyi Dong[2,3,†], Pan Zhang[2], Bin Wang[2], Conghui He[2], Jiaqi Wang[2], Dahua Lin[2], Weiming Zhang[1,†], Nenghai Yu[1]

[1]Anhui Province Key Laboratory of Digital Security, University of Science and Technology of China

[2]Shanghai AI Laboratory [3]The Chinese University of Hong Kong

{hqd0037@mail., zhangwm@, ynh@}ustc.edu.cn {xydong@, dhlin@}ie.cuhk.edu.hk

{zhangpan@, wangbin@, heconghui@}pjlab.org.cn wjqdev@gmail.com

1. Limitation & Social Impact

In this section, we clarify the weaknesses of our proposed OPERA and the potential social impact incurred by it.

Limitations. We have identified two main limitations of the proposed approach: 1) The first limitation lies in it can not address all kinds of the hallucinations of MLLMs. It is understandable since our approach serves as a nearly free lunch method for MLLMs without incurring additional costs. Upon reviewing the failure cases of OPERA, we discern various causes for hallucinated content. One likely reason is MLLMs' strong biases in the generated content. The knowledge aggregation mechanism of MLLMs causes subsequent token generation to overly rely on summary tokens while neglecting detailed information from the front-most image tokens. For instance, MLLMs may easily hallucinate "cars" in subsequent tokens when the preceding content mentions "road". Such hallucinations should blame MLLM's strong bias between "road" and "cars", which is learned during the training phase. In this scenario, OPERA can well handle many cases unless the model's bias is too strong that it is challenging to find a suitable candidate during the retrospection-reallocation phase. Another probable reason is that MLLMs' visual perception is not sufficiently robust. MLLMs can be misled by similar shapes, colors of objects, or issues related to low resolution. In these cases, OPERA faces challenges, constrained by the model's visual capabilities. 2) The second limitation is that, OPERA demonstrates marginal gains when addressing hallucinations in short answers (< 10 tokens), primarily due to the hysteresis of knowledge aggregation patterns. OPERA excels in handling hallucinations occurring in long sequences. To overcome this limitation, a potential solution is to enhance the metric for detecting knowledge aggregation patterns and increase its sensitivity.

Social impacts. There is no potential for social harm caused by OPERA. Instead, it holds the promise to significantly propel the advancement of MLLMs. OPERA serves as an inspiration for the community to delve into more effective approaches for alleviating MLLMs' hallucination issue without incurring additional costs. Such approaches can better generalize on different kinds of MLLMs.

2. Ablation Study on Hyper-parameters

In this section, we give detailed ablation studies for hyper-parameters, including two key components, the number of candidates N_{can}, the scale factor σ, the penality weight α, and the threshold r of retrospection. Despite the best parameter of different MLLMs are a little bit different, OPERA is generally robust on the varying settings of hyper-parameters and outperforms the baselines. In our paper, we simply adopt a default setting with $N_{can} = 5$, $\sigma = 50$, $\alpha = 1$, and $r = 15$ for all MLLMs.

Key components. Here we ablate the two components proposed in OPERA, *i.e.*, the over-trust penalty and the retrospection-reallocation strategy. As the results shown in Table 2, when we discard both components, our method degrade to standard Beam search and presents worst perfoemance. Equipped either of the two components can help MLLM models hallucinate less, where the over-trust penalty contributes relatively more to the final performance. It is promising, since not all of generated sequences need to retrospect during decoding, unless encountering the knowledge aggregation patterns.

Number of candidates N_{can}. To prevent the model give unreasonable output, we restrict the prediction of each beam within the top-N_{can} highest vocabularies in the logit. Note

*Work done during an internship in Shanghai AI Laboratory.

†Corresponding authors.

Setting	N_{can}	σ	α	r	InstructBLIP		MiniGPT-4		LLaVA-1.5		Shikra	
					C_S	C_I	C_S	C_I	C_S	C_I	C_S	C_I
Beam Search	-	-	-	-	55.6	15.8	30.6	9.5	48.8	13.9	50.4	13.3
A1	2	50	1	15	**43.8**	**13.1**	29.8	10.8	**41.2**	**12.0**	43.0	12.8
A2	3	50	1	15	46.4	13.2	30.0	10.0	43.8	12.8	39.4	12.7
A3	5	50	1	15	46.4	14.2	**26.2**	**9.5**	44.6	12.8	36.2	12.1
A4	8	50	1	15	49.6	14.6	29.0	10.1	49.0	13.4	**33.3**	**11.5**
A5	10	50	1	15	51.4	15.0	30.4	10.0	48.0	13.2	34.4	11.6
B1	5	40	1	15	47.6	14.3	27.8	10.2	46.9	13.3	45.4	12.8
B2	5	45	1	15	47.2	14.5	26.8	9.8	47.8	13.3	41.2	12.3
B3	5	50	1	15	46.4	14.2	26.2	9.5	**44.6**	**12.8**	36.2	12.1
B4	5	55	1	15	44.2	**13.9**	**25.6**	**9.2**	47.5	13.3	35.4	11.7
B5	5	60	1	15	**44.0**	14.3	26.6	10.9	44.5	13.0	**33.8**	**11.7**
C1	5	50	0.1	15	47.6	14.4	26.6	9.7	46.4	12.8	40.2	12.4
C2	5	50	0.5	15	46.2	14.3	27.6	9.7	46.4	13.3	35.6	**11.5**
C3	5	50	1	15	46.4	14.2	**26.2**	**9.5**	**44.6**	**12.8**	36.2	12.1
C4	5	50	5	15	46.0	**13.8**	27.2	9.9	47.6	13.5	39.2	13.2
C5	5	50	10	15	**45.4**	14.0	26.4	9.5	46.4	13.2	**35.4**	12.6
D1	5	50	1	5	52.0	14.8	**24.9**	9.8	45.0	13.0	40.2	12.7
D2	5	50	1	10	50.4	14.8	26.4	10.1	45.3	12.9	**36.0**	**11.5**
D3	5	50	1	15	46.4	14.2	26.2	**9.5**	**44.6**	**12.8**	36.2	12.1
D4	5	50	1	20	42.6	13.4	27.1	9.7	45.6	13.0	37.0	12.1
D5	5	50	1	25	**41.8**	**13.1**	27.6	9.8	45.0	12.9	40.0	13.3

Table 1. Ablation studies on the hyper-parameters used in our OPERA, including the number of candidates N_{can}, the scale factor σ, the penalty weight α and the rollback threshold r. Denote CHAIR$_S$ as C_S and CHAIR$_I$ as C_I. Lower values mean less hallucinations.

Setup	P	R	InstructBLIP		MiniGPT-4		LLaVA-1.5		Shikra	
			C_S	C_I	C_S	C_I	C_S	C_I	C_S	C_I
A	✗	✗	55.6	15.8	30.6	9.5	48.8	13.9	50.4	13.3
B	✗	✓	50.0	14.6	27.3	10.1	46.4	12.9	46.8	13.0
C	✓	✗	48.2	**13.8**	27.4	10.0	45.2	13.0	41.8	13.9
D	✓	✓	**46.4**	14.2	**26.2**	**9.5**	**44.6**	**12.8**	**36.2**	**12.1**

Table 2. Ablation results on two components. "P" denotes the over-trust penalty, "R" denotes retrospection-reallocation strategy.

that N_{can} is a configurable parameter like N_{beam} in Beam Search [1, 4, 7]. An appropriate setup of N_{can} can greatly improve the performance of OPERA. Too small N_{can} may decrease the effect of retrospection-reallocation, while too large N_{can} probably engages some unreasonable vocabularies that are irrelevant with the whole sequence. The results are listed in Table 1. InstructBLIP [3] and LLaVA-1.5 [6] may prefer smaller N_{can}, while MiniGPT-4 [10] prefers $N_{can} = 5$ and Shikra [2] prefers larger N_{can}.

Scale Factor σ. Before depicting the knowledge aggregation pattern through column-wise multiplication in attention maps, we set a scale factor σ to scale up attention values which are usually too small. As the results presented in Table 1, different MLLM models prefer different scale factors, probably because the varying sequence lengths (*e.g.*, LLaVA-1.5-7B has 576 image tokens while MiniGPT-4-7B has only 32 image tokens) result in different magnitudes of self-attention weight values (Note that the sum of self-attention weights should be 1). In other words, σ is a configurable parameter for users to pursue the best performance of their own MLLM model in the rough range of 40 to 60. For simplicity, we set σ as 50, a balanced choice that performs not bad on different MLLMs.

Penalty weight α. We further ablate the weight of the introduced penalty term that is incorporated with the model logit. From the results in Table 1, we can observe that OPERA's performance is relatively robust when α varies. Different MLLMs may prefer different α, but the numerical fluctuations are generally slight. For simplicity, we unify α as 1 for different MLLMs.

Rollback threshold r. We consider the location overlap of the maximum column-wise scores of several consecutive tokens as the condition of retrospection, where we set a threshold r for the count of overlap. If the count of overlap reaches the threshold r, the rollback will be triggered. Consequently, the choice of r seems crucial and a ablation study is necessary. The abaltion results are shown in Table 1. We can observe that InstructBLIP shows less hallucinations when $r = 25$ while the other three MLLMs show have the better perofrmance when $r = 15$. Therefore, we assign r as 15 by default.

GPT-4 Prompt

Please help me judge if the comment of this image is hallucination or correct.
I will give you a list of region description of a image. The format is [x1, y1, x2, y2]: region description, where [x1, y1, x2, y2] is the bounding box of the region. Highly overlapping bounding boxes may refer to the same object. This is the ground truth information of the image. Your judgement should base on this information. However, this information only describe the objects in the region of image, so it cannot describe the subjective part of the image, e.g., atmosphere, style, emotion. In that case, you can return “Cannot judge”.
Also, I will give you a list of comments of the image for you to judge if it is hallucination. Please give a judgement one by one along with the reason.
You should pay extra attention to the hallucination, which refers to the part of comments that are inconsistent with the descriptions, specially claiming the existence of something not present in the descriptions.

If a comment is hallucination, please help me rewrite it. When rewrite the comment, sound like you are looking at the image directly.
Each rewritten comments should compose a description about the image which is correct, detailed, smooth and has strong readability.
If not hallucination (correct or cannot judge), keep the original comment.

Your output should be:
Judgement:
1. hallucination or correct or cannot judge: <reason>
2. ...
Revised Sentences:
1. ...
2. ...

Here are the region descriptions of the image:
{}
Here is the comment for you to judge if it is hallucination and revise:
{}

Table 3. The prompt used for GPT-4 evaluation.

GPT-4V(ision) Prompt

You are required to score the performance of two AI assistants in describing a given image. You should pay extra attention to the hallucination, which refers to the part of descriptions that are inconsistent with the image content, such as claiming the existence of something not present in the image or describing incorrectly in terms of the counts, positions, or colors of objects in the image. Please rate the responses of the assistants on a scale of 1 to 10, where a higher score indicates better performance, according to the following criteria:
1: Accuracy: whether the response is accurate with respect to the image content. Responses with fewer hallucinations should be given higher scores.
2: Detailedness: whether the response is rich in necessary details. Note that hallucinated descriptions should not countas necessary details.
Please output the scores for each criterion, containing only two values indicating the scores for Assistant 1 and 2, respectively. The two scores are separated by a space. Following the scores, please provide an explanation of your evaluation, avoiding any potential bias and ensuring that the order in which the responses were presented does not affect your judgment.

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

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

Output format:
Accuracy: <Scores of the two answers>
Reason:

Detailedness: <Scores of the two answers>
Reason:

Table 4. The prompt used for GPT-4V(ision) evaluation.

3. Details of GPT-4 Evaluation

We generally follow the GPT-4 evaluation proposed in HalluBench [9] and implement it on VG dataset. Each image in VG [5] dataset has the detailed ground-truth descriptions about all of the appearing objects. Since GPT-4 is not able to deal with image data, we integrate all of ground-truth descriptions into the input prompt to help GPT-4 comprehend the image content. Then, given the MLLM's generated description on the image with "`Please describe this image in detail.`", GPT-4 are required to judge whether each sentences of MLLM's description has hallucinated contents. This evaluation is quite strict, where GPT-4 judges any MLLM's descriptions as hallucinations if they are deviated from the ground-truth descriptions in terms of quantity, color, location, activity, or direction.

Metrics. There are six metrics considered, which include:

- *The number of sentences per image (SPI).* It reflects the detailedness of MLLM's description at the sentence level.
- *The number of words per image (WPI).* It reflects the detailedness of MLLM's description at the word level.
- *The number of hallucinated sentences per image (HSPI).* It reveals the hallucination degree of MLLM's description at the sentence level. Any sentences that contain hallucinated contents are taken into calculation.
- *The number of hallucinated words per image (HWPI).* It reveals the hallucination degree of MLLM's description at the word level. Any words related with hallucinated contents are taken into calculation.
- *The ratio of hallucinated sentences (HSR).* The average ratio of hallucinated sentences in all sentences of MLLM's descriptions on different images.
- *The ratio of hallucinated words (HWR).* The average ratio of hallucinated words in all words of MLLM's descriptions on different images.

Prompt. As shown in Table 3, our adopted GPT-4 prompt is generally based on HalluBench [9].

4. Details of GPT-4v Evaluation

Following [8], we conduct the dual evaluation on GPT-4V(ision) for Beam search and our proposed OPERA. Given a trained MLLM model and a image, we respectively use Beam search decoding and OPERA decoding to obtain two descriptions with the prompt "`Please describe this image in detail.`". Then, we adopt the prompt shown in Table 4 to ask GPT-4V to rate the two description based on the image on a scale of 0 to 10, where the rating involves two aspects, *i.e.*, Accuracy and Detailedness. The accuracy reflects the consistency between the description and the given image. If GPT-4V thinks any content in this description is inconsistent with the given image, namely higher hallucinations, it will get lower score. The detailedness reflects the degree of expres-

Input image:

Instruction: *Please describe this image in detail.*

InstructBLIP (Beam Search):
The image features a large white polar bear sitting in the water, holding a carrot in its mouth. The bear appears to be relaxing and enjoying its time in the water. The carrot is positioned close to the bear's mouth, suggesting that it might be part of the bear's meal or snack. ***There are several smaller carrots scattered around the scene, possibly representing other parts of the bear's diet. In addition to the carrots, there are various other objects visible in the water, such as leaves, sticks, and small rocks.*** *Overall, the scene depicts a peaceful moment for the polar bear as it enjoys its aquatic surroundings.*

InstructBLIP (OPERA):
The image depicts a polar bear sitting in the water, holding a carrot in its mouth. The bear is positioned in the middle of the water, surrounded by various sized waves and ripples. The carrot appears to be a favorite treat for the polar bear, as it holds it tightly in its mouth. The scene captivates the viewer with the serene and peaceful atmosphere of the polar bear enjoying its meal in the water.

Figure 1. OPERA's performance on reducing hallucinations.

sive ability, *i.e.*, how comprehensive does the description characterize the image.

The prompt adopted for GPT-4V is listed in Table 4. It requires GPT-4V to ignore the bias incurred by the sequntial order and pay extra attention to the objects mentioned by MLLM's descriptions but not appear in the image, including incorrect colors, positions, or relationships. GPT-4V comprehensively analyzes MLLM's description, using its strong abilities that are closed to human.

5. Potentials for Eliminating Repetition

Repetition is also a problem of MLLMs, usually manifested as the model's incessantly repeating on the particular sentence. We notice that OPERA can well handle such repetition, as showcased in Figure 2. Interestingly, the self-attention map of repeated sentences appears periodic knowledge aggregation patterns. Accordingly, OPERA can help the sequence to retrospect and reallocate at other appropriate vocabularies like "eos" token.

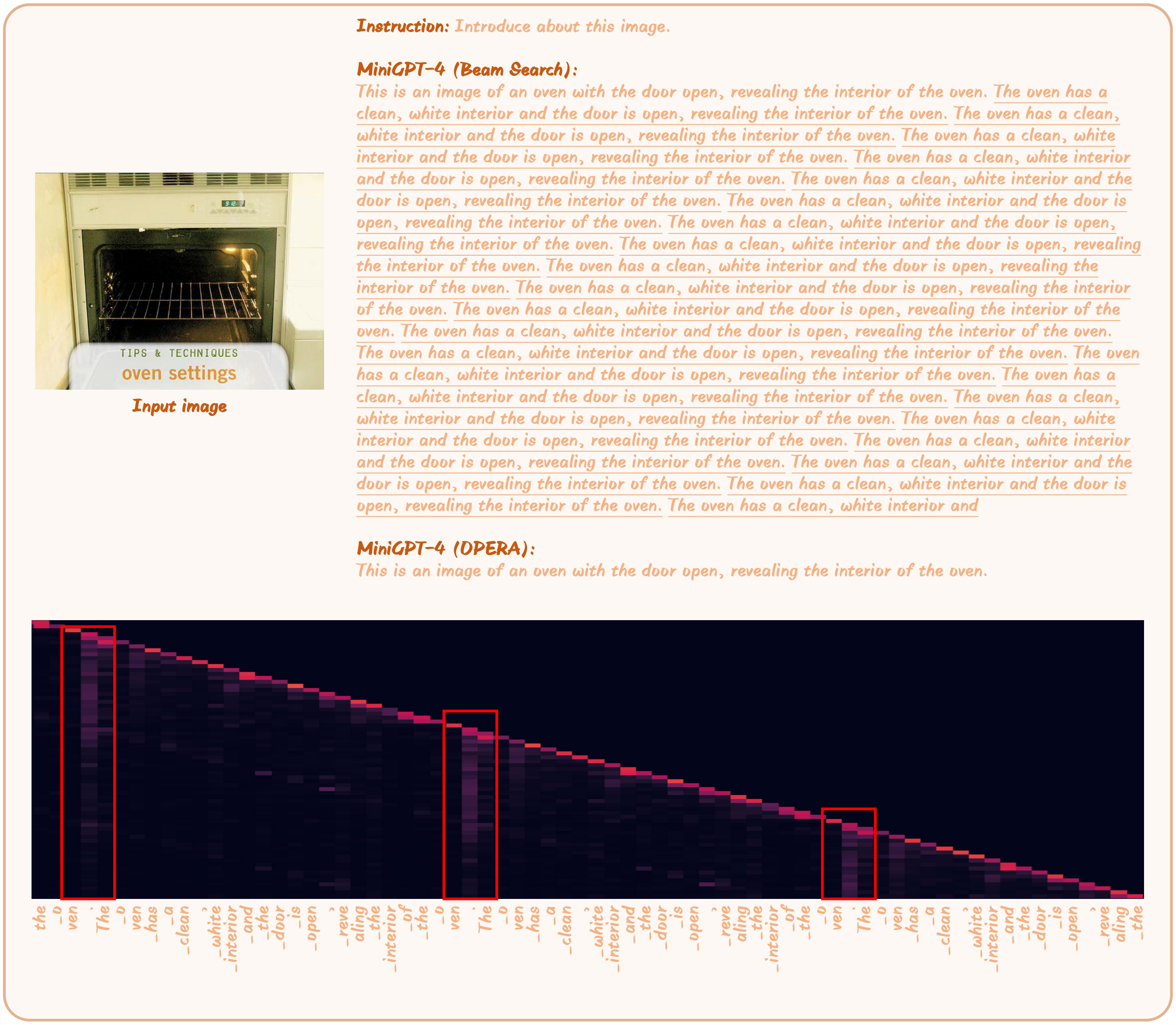


Figure 2. OPERA's potentials for eliminating repetition.

6. Qualitative Results

We provide several cases that proves OPERA's strong ability on mitigating hallucinations. These cases uses various MLLMs and different instructions including "`Please describe this image in detail.`", "`What can you see in this image?`", and "`Introduce about this image.`". The cases are shown in Figure 1, Figure 3, Figure 4 and Figure 5 (Please check the next pages).

References

[1] Nicolas Boulanger-Lewandowski, Yoshua Bengio, and Pascal Vincent. Audio chord recognition with recurrent neural networks. In *ISMIR*, pages 335–340. Curitiba, 2013. 2

[2] Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic. *arXiv preprint arXiv:2306.15195*, 2023. 2

[3] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat 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] Alex Graves. Sequence transduction with recurrent neural networks. *arXiv preprint arXiv:1211.3711*, 2012. 2

[5] Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. Visual genome:

Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73, 2017. 4

[6] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning, 2023. 2

[7] Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. *Advances in neural information processing systems*, 27, 2014. 2

[8] Shukang Yin, Chaoyou Fu, Sirui Zhao, Tong Xu, Hao Wang, Dianbo Sui, Yunhang Shen, Ke Li, Xing Sun, and Enhong Chen. Woodpecker: Hallucination correction for multimodal large language models. *arXiv preprint arXiv:2310.16045*, 2023. 4

[9] Zhiyuan Zhao, Bin Wang, Linke Ouyang, Xiaoyi Dong, Jiaqi Wang, and Conghui He. Beyond hallucinations: Enhancing lvlms through hallucination-aware direct preference optimization. *arXiv preprint arXiv:2311.16839*, 2023. 4

[10] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *arXiv preprint arXiv:2304.10592*, 2023. 2

Input image

Instruction: Please describe this image in detail.

LLaVA-1.5 (Beam Search):
The image captures a man skillfully riding a wave on a surfboard in the ocean. He is wearing a wetsuit and appears to be enjoying the thrill of surfing. The surfer is positioned in the center of the scene, with the surfboard beneath him as he skillfully navigates the wave.

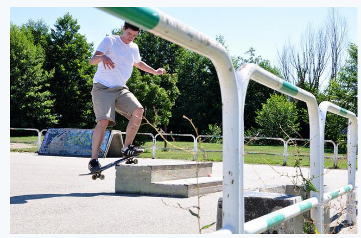
In the background, there are a few other people in the water, possibly watching the surfer or waiting for their turn to ride the waves. Overall, the scene showcases the excitement and challenge of surfing in the ocean.

LLaVA-1.5 (OPERA):
The image captures a man skillfully riding a wave on a surfboard in the ocean. He is wearing a wetsuit and appears to be enjoying the thrill of surfing. The surfer is positioned in the center of the scene, with the surfboard beneath him as he skillfully navigates the wave.The ocean is filled with waves, creating a dynamic and exciting environment for the surfer. The waves vary in size and shape, adding to the challenge and excitement of the surfing experience.

Instruction: Please describe this image in detail.

LLaVA-1.5 (Beam Search):
The image captures a young man skillfully performing a trick on his skateboard at a skate park. He is in mid-air, jumping over a cement bench while riding the skateboard. The skateboarder is the main focus of the scene, showcasing his talent and athleticism. The skate park is surrounded by trees, providing a pleasant and natural environment for the skateboarder to practice his tricks. There are a few other people in the background, possibly watching the skateboarder or waiting for their turn to perform their own tricks.

LLaVA-1.5 (OPERA):
The image captures a young man skillfully performing a trick on his skateboard at a skate park. He is riding the skateboard up the side of a cement ramp, showcasing his talent and balance. The skateboarder is the main focus of the scene, with his skateboard prominently visible beneath him. The skate park is surrounded by a lush green field, creating a pleasant and inviting atmosphere for skateboarders and spectators alike.



Input image

Input image

Instruction: What can you see in this image?

LLaVA-1.5 (Beam Search):
The image features a street corner with two blue and white street signs pointing in opposite directions. One of the signs reads "Richmond Library," while the other reads "West Market St." The signs are mounted on a metal pole, and they are positioned in front of a building.

In addition to the street signs, there are two traffic lights visible in the scene. One traffic light is located on the left side of the image, while the other is on the right side. The presence of these traffic lights suggests that the area is well-regulated for vehicular and pedestrian traffic.

LLaVA-1.5 (OPERA):
The image features a street corner with two blue and white street signs pointing in opposite directions. The signs indicate the names of the streets, "Richmond Library" and "West Market Street." The street signs are positioned on a pole, and the sky can be seen in the background, providing a clear view of the surroundings.

Instruction: Introduce about this image.

LLaVA-1.5 (Beam Search):
The image features a red stop sign prominently placed on a wooden dock or pier. The stop sign is positioned in the foreground of the scene, making it the focal point of the image.

In the background, there are several people scattered around the area, possibly enjoying the view or engaging in various activities. Additionally, there are two benches visible in the scene, providing a place for people to sit and relax while taking in the surroundings.

LLaVA-1.5 (OPERA):
The image features a red stop sign prominently placed on a wooden dock or pier. The stop sign is positioned in the foreground, making it the focal point of the scene. In the background, there is a building visible, adding to the overall atmosphere of the location.

Input image

Figure 3. OPERA's performance on reducing hallucinations of LLaVA-1.5-7B.

Instruction: What can you see in this image?

InstructBLIP (Beam Search):
The image features a group of three cows walking down a street in a residential area. The cows are positioned in the middle of the street, with one cow on the left side and two cows on the right side. They seem to be following each other as they make their way down the road. There are several buildings visible in the background, including a house on the left side of the street and another house on the right side of the street. **A car is parked on the left side of the street, near the first house, and another car is parked on the right side of the street, closer to the second house.**

InstructBLIP (OPERA):
The image features a group of three cows walking down a street in a residential area. The cows are positioned in the middle of the street, with one cow on the left side and two cows on the right side. They seem to be following each other as they make their way down the street. There are several buildings visible in the background, including a house on the left side of the image and another building on the right side. A fence can also be seen in the scene, likely separating the cows from the residential area. Overall, the image captures a peaceful moment as the cows walk down the street, surrounded by the residential surroundings.

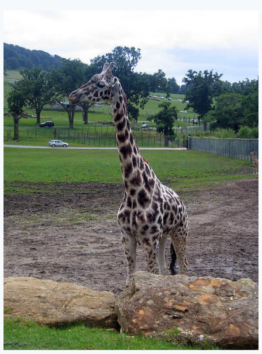
Input image

Instruction: What can you see in this image?

InstructBLIP (Beam Search):
The image features a woman standing next to a giraffe, feeding it with her hand. She is wearing a hat and appears to be interacting with the giraffe in a zoo setting. The giraffe's head is positioned close to the woman's hand as she feeds it. There are several trees visible in the background, providing a natural setting for the interaction between the woman and the giraffe. **In addition to the woman and the giraffe, there are two other people in the scene. One person is located on the left side of the image, while the other person is situated on the right side. Both individuals appear to be observing the interaction between the woman and the giraffe.**

InstructBLIP (OPERA):
The image features a woman standing next to a giraffe, feeding it with her hand. The giraffe is positioned in the center of the scene, while the woman is on the left side of the image. She is wearing a hat and appears to be engrossed in feeding the giraffe. There are several trees visible in the background, providing a natural setting for the encounter between the woman and the giraffe.

Input image



Input image

Instruction: Introduce about this image.

InstructBLIP (Beam Search):
The image features a giraffe standing in a grassy field, surrounded by trees and rocks. The giraffe is the main focus of the scene, with its long neck stretching out towards the left side of the image. There are several cars parked in the background, likely belonging to visitors who have come to see the giraffe. **In addition to the giraffe, there is another giraffe visible in the right side of the image, but it is not as prominent as the main giraffe.**

InstructBLIP (OPERA):
The image features a giraffe standing in a grassy field, surrounded by rocks and trees. The giraffe is the main focus of the scene, with its head and neck stretched out towards the left side of the image. There are several cars parked in the background, likely belonging to visitors to the zoo or wildlife park. The giraffe appears to have a curious expression on its face, possibly enjoying the scenery around it.

Instruction: Please describe this image in detail.

InstructBLIP (Beam Search):
The image features a dining table with a plate of chocolate cake on it. The cake is decorated with white frosting, making it visually appealing and delicious-looking. There are several pieces of cake on the plate, arranged in various sizes and shapes. Some of the cake pieces are closer to the edge of the plate, while others are more centrally located. **In addition to the cake, there is a bowl placed on the table, possibly containing additional desserts or condiments. A wine glass is also visible on the table, adding to the festive atmosphere of the scene.**

InstructBLIP (OPERA):
The image features a dining table with a plate of chocolate cake on it. The cake is decorated with white frosting and drizzles, making it look like a delicious dessert. There are multiple slices of the cake on the plate, with some slices closer to the edge and others closer to the center. The cake appears to be a Christmas-themed dessert, possibly served during a holiday celebration.

Input image

Figure 4. OPERA's performance on reducing hallucinations of InstructBLIP-7B.

Instruction: Please describe this image in detail.

Input image

Shikra (Beam Search):
The image features a white vase filled with yellow flowers. The vase is placed in the center of the scene, and the flowers are arranged in various positions within the vase, creating a beautiful and lively arrangement.

In addition to the main vase, there are **three** smaller vases around it. Two of the smaller vases are placed to the right of the main vase, **while the third smaller vase is placed to the left of the main vase.** The combination of the larger vase and multiple smaller vases creates a visually pleasing and harmonious composition.

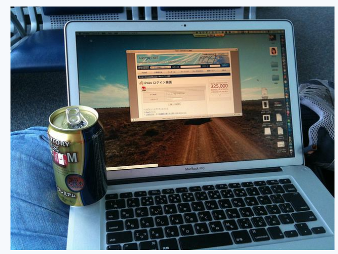
Shikra (OPERA):
The image features a white vase filled with yellow flowers, placed on a white table. There are three vibrant flowers in the vase, each with a different height and position. The flowers are arranged in such a way that they create a lively and colorful focal spot, adding a touch of beauty to the scene.

Instruction: Please describe this image in detail.

Shikra (Beam Search):
The image features a laptop computer sitting open on a table. A can of soda is placed next to the laptop, creating a casual atmosphere. On the screen of the laptop, there is a window open, **possibly displaying a Facebook page**.

Additionally, there is a bird visible in the background, adding a touch of nature to the scene. The overall setting appears to be a relaxed environment where someone might be working, studying, or enjoying leisure time using the laptop and soda.

Shikra (OPERA):
The image features a laptop computer sitting open on a table. Near this laptop, there is a can of soda, possibly a can of beer, placed on top of the keyboard. The beverage can is in the foreground, occupying a larger portion and contrasting with the laptop in the background.



Input image

Instruction: Please describe this image in detail.

Input image

Shikra (Beam Search):
The image features a young child, likely a baby, sitting in a high chair in front of a birthday cake. The child appears to be enjoying the cake, and the cake is placed in the middle of the high chair, within reach.

A knife is visible on the left side of the image, possibly used for cutting the cake. The room is furnished with a dining table positioned in the background, and a chair situated on the left side of the high chair.

Shikra (OPERA):
The image features a young child sitting in a high chair in front of a birthday cake. The child appears to be curious about the cake, possibly trying to eat or touch it. The cake is placed in the middle of the high chair and seems to be the main focus of the scene.

Instruction: What can you see in this image?

Shikra (Beam Search):
In this image, I can see a beautiful valley filled with lush green grass. There are several horses grazing in the meadow, enjoying the serene and picturesque surroundings. The valley is surrounded by tall trees, adding to the natural beauty of the scene.

Additionally, a dining table and a couple of chairs can be spotted near the edge of the valley, providing a place for people to sit, relax, and take in the view. The overall atmosphere of the scene is peaceful and idyllic.

Shikra (OPERA):
In this image, you can see a beautiful valley filled with lush green grass and a river flowing through a lush green field. There are several horses grazing on this picturesque countryside, enjoying the serene environment. The valley is surrounded by majestic mountains, adding to the stunning beauty of the landscape.

Input image

Figure 5. OPERA's performance on reducing hallucinations of Shikra-7B.