SHIFT: Smoothing Hallucinations by Information Flow Tuning for Multimodal Large Language Models
— Supplementary Material —

Sudong Wang[1,2], Yunjian Zhang[3]*, Yao Zhu[3]*, Enci Liu[4]
Jianing Li[3], Yanwei Liu[1], Xiangyang Ji[3]
[1]Institute of Information Engeering, Chinese Academy of Sciences;
[2]Nanyang Technological University; [3]Tsinghua University; [4]Columbia University
SWANG049@e.ntu.edu.sg, sdtczyj@gmail.com, ee_zhuy@zju.edu.cn

1. Experimental Settings

In our experiments, we set $\epsilon = 0.5$, $\delta = 0.5$, and $\alpha = 0.9$ for all benchmarks. The prompt we use for the descriptions experiment is "Please describe this image in detail". The models we use have the following Huggingface identifiers:

- liuhaotian/llava-v1.5-7b
- MAGAer13/mplug-owl2-llama2-7b
- Salesforce/instructblip-vicuna-7b

2. Ablation Study on Hyper-parameters

In our approach, the tuning parameter α is important for balancing the continuous knowledge from the preceding layers and the injected knowledge from the mutation layers. In this section, we give a detailed ablation study for this parameter, and the results are shown in Table 1. It can be observed that as α increases, the overall hallucination tends to diminish, indicating that reducing the influence of injected information can help prevent hallucinations. Although larger α will make the model's output more faithful to the input features. However, deeper layers generally focus on global information, as a result,overly strong smoothing may lead to worse fluency. For example, in our experiments, we find that when $\alpha = 1$, although hallucinations are reduced, the readability of the output text is compromised, suggesting that the injected information also contains knowledge relevant to textual fluency. Considering these factors, an alpha of 0.9 achieves an satisfying balance between factual accuracy and text quality.

Table 1. Results of the ablation study on α.

Decoding	α	LLaVA-1.5 [4]		mPLUG-Owl2 [6]	
		CHAIR_S	CHAIR_I	CHAIR_S	CHAIR_I
Greedy	0.1	52.8	13.8	59.4	15.5
	0.2	52.4	14.3	57.8	15.9
	0.3	52.6	13.9	57.4	15.7
	0.4	50.2	13.3	56.4	15.9
	0.5	51.0	12.8	56.2	15.3
	0.6	47.4	**12.1**	55.8	15.5
	0.7	46.8	12.6	53.9	15.4
	0.8	46.6	12.3	52.2	15.3
	0.9	**43.8**	12.4	**50.2**	**15.2**
Beam Search	0.1	46.8	13.2	58.5	15.3
	0.2	46.7	13.5	57.6	15.2
	0.3	46.0	13.1	57.7	15.5
	0.4	44.9	12.7	56.1	14.8
	0.5	42.1	12.2	55.0	14.2
	0.6	41.5	11.8	52.3	14.0
	0.7	40.2	11.3	50.2	13.8
	0.8	38.6	11.5	49.3	13.9
	0.9	**36.7**	**10.5**	**47.4**	**13.7**
Nucleus	0.1	51.6	14.4	60.4	16.0
	0.2	52.0	14.2	60.2	16.1
	0.3	50.8	13.3	58.7	16.8
	0.4	49.8	13.5	57.9	16.4
	0.5	48.2	13.1	56.2	16.2
	0.6	47.2	12.5	55.1	16.2
	0.7	45.8	12.9	52.5	15.8
	0.8	44.4	13.1	49.8	15.1
	0.9	**42.0**	**11.6**	**47.2**	**14.4**

3. Effect of SHIFT when MLLM Scales Up

Our evaluation extends to LLaVA-1.5-13b [4], assessing the scalability of SHIFT across different MLLM magnitudes. The results on the CHAIR benchmark [5] and POPE benchmark [3] are shown in Table 2 and Table 3, and the greedy decoding strategy is used. It can be observed that the performance on the hallucination evaluation benchmark is similar before and after scaling up the model, indicating that increasing the model's parameters does not fundamentally resolve the hallucination issue. This further highlights the challenge and necessity of addressing hallucinations in MLLMs. Compared to the vanilla model, SHIFT consistently improves model performance across all settings, demonstrating its robustness across different models.

Table 2. Results on the CHAIR benchmark, where a smaller number indicates less hallucinations. LLaVA-1.5-13b is used.

Method	CHAIR$_S$	CHAIR$_I$
Vanilla	53.6	14.6
SHIFT	**47.0**	**11.7**

Table 3. Results on the POPE benchmark. The best results are in **bold**. LLaVA-1.5-13b is used.

Dataset	**Setting**	**Method**	Acc	F1
MSCOCO	Random	Vanilla	88.37	88.98
		SHIFT	**88.67**	**89.21**
	Popular	Vanilla	85.13	86.34
		SHIFT	**85.57**	**86.67**
	Adversarial	Vanilla	79.07	81.79
		SHIFT	**79.60**	**82.14**
GQA	Random	Vanilla	83.46	85.68
		SHIFT	**83.93**	**86.02**
	Popular	Vanilla	75.40	80.08
		SHIFT	**75.82**	**80.31**
	Adversarial	Vanilla	68.60	75.91
		SHIFT	**69.03**	**76.10**

4. Extend to Large Language Models

Based on our observations, we believe that the information injected in the mutation layer may not originate from the input image. Instead, it is more likely to come from the prior knowledge acquired by the language model during pretraining. When this knowledge conflicts with the input image, it may trigger hallucinations. To further validate this hypothesis and explore the influence of the language model's prior knowledge on token prediction, we conduct an additional experiment on the LLaVA-1.5 model [4]. In this experiment, we intentionally provide incorrect knowledge as prompts and guided the LLM to answer questions based on this misinformation. We use an example to illustrate our analyses, with the prompt and the results are shown in Table 4 and Table 5, respectively. As can be seen, even though we provide incorrect information to the vanilla model, it still manages to output the correct answer, indicating that its token predictions are not entirely dependent on the input; its prior knowledge acquired through pretraining also plays a vital role in the generation process. However,

Table 4. The prompt used for evaluating LLMs.

LLM Prompt
You are an AI assistant. Use only the information provided below to answer the questions. [Knowledge Base:] The signing of the Treaty of Versailles occurred in Tokyo, Japan in 1919. [Question:] Where was the Treaty of Versailles signed in 1919?

when we apply SHIFT near the output layer, the model produce answers entirely based on our prompts, even when the prompts contain incorrect information. This suggests that, in this case, the internal priors have minimal influence on token prediction. This example demonstrates that applying SHIFT can reduce the impact of the model's internal priors on the generation process, making it more reliant on input information. Considering that hallucinations in MLLMs mainly arise from outputs not being faithful to inputs, applying SHIFT can help ensure that the model's output more closely aligns with the input description. This experiment explains the principle of SHIFT from the perspective of the language model.

Table 5. LLM evaluation results.

Model	Vanilla	SHIFT on the 30-th layer
Answer	Paris, France	**Tokyo, Japan**

5. Details of GPT-4v Evaluation

Following [2], we conduct the GPT-4v Evaluation for our proposed SHIFT on the greedy decoding. For the LLaVA-1.5 model and an image, we respectively use vanilla greedy and SHIFT to obtain two descriptions with the prompt "Please describe this image in detail". Then, we adopt the prompt shown in Table 6 to ask GPT-4v to rate the two description based on the image on a scale of 0 to 10. This evaluation comprehensively analyzes MLLM's description from a human-like perspective. An illustrative evaluation example is presented in Figure 1.

6. More Cases on the Information Flow

In this section, we provide additional examples of the information flow in MLLMs to further illustrate the prevalence of the mutation phenomenon. In the experiments, given an

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

input image, we use "Please describe this image in detail" to prompt the MLLM to generate a relevant description. Figure 2 shows four hallucinated examples and their corresponding Jensen-Shannon divergences (JSDs). It can be observed that for all hallucinated words, including "a black and white shirt", "shirt", "bowls", and "soccer", their divergences show clear mutations near the output layers. This phenomenon is even more pronounced in the first token associated with the hallucination. This occurs because, during the next-token prediction process, the model is directly injected with hallucinated knowledge that does not originate from the input, altering the token prediction and subsequently affecting the following autoregressive process, ultimately leading to hallucinations like "a black and white shirt" that involve multiple words.

We then analyze the probability changes of some hallucinated tokens across the intermediate layers in the network, with the results shown in Figure 3. For the hallucinated tokens, such as "car", "people", "boats", and "person", their probabilities undergo mutations that make them become the dominant tokens, persisting until the output and are ultimately used by the model to describe the image. In contrast, before the mutations, the dominant tokens are the correct ones. These examples further illustrate the feasibility of processing the mutation layers to eliminate hallucinations.

7. Qualitative Results

We provide several examples for proving the capability of SHIFT 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". Additionally, we also present some examples on the VQA task. The results are shown in Figure 4, Figure 5, and Figure 6.

References

[1] Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. MME: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

[2] Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. OPERA: alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13418–13427, 2024. 2

[3] Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 292–305, 2023. 1

[4] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. 1, 2

[5] Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2018. 1

[6] Qinghao Ye, Haiyang Xu, Jiabo Ye, Mingshi Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owi2: Revolutionizing multi-modal large language model with modality collaboration. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 13040–13051, 2023. 1

Input:

[Assistant 1]

The image depicts a woman standing on a tennis court, holding a tennis racquet and about to hit a tennis ball. She is wearing white kilt and a white top, which is a common color scheme for tennis players. The woman is positioned near the center of the court, ready to make contact with the ball. In addition, a green tennis ball, is visible in the air, as if it has been hit by the woman. The scene captures the excitement and action of a game of tennis.

[Assistant 2]

The image features a woman playing tennis on a court, holding a tennis racket and preparing to hit a tennis ball. She is standing on the court, focused on the game. The tennis ball is in the air, close to the woman, and she is ready to make contact with it. There are two other people in the scene, one standing near the left side of the court and another person on the right side. A car can be seen in the background, parked behind the tennis court.

GPT4-v Output:

Accuracy:
Scores of the two answers:
Assistant 1: 9
Assistant 2: 4
Reason: Assistant 1 provides an accurate description of the visible contents of the image with marginal errors or assumptions. The woman, her attire, and her actions are described precisely as they appear. The hallucination here is minimal to none, as it accurately captures the scenario. Assistant 2, however, introduces significant inaccuracies: the mention of two other people and a parked car none of which are visible in the image. This constitutes a clear example of hallucination, thus lowering the accuracy score significantly.

Detailedness:
Scores of the two answers:
Assistant 1: 8
Assistant 2: 5
Reason: Assistant 1 offers a detailed account of the scene inclusive of the tennis attire, explicit action of playing tennis, and even the movement of the tennis ball, adding to a rich description of the scenario. Assistant 2, while initially correct about the tennis player, dilutes the quality of the description with unnecessary and incorrect details about other people and a car, which do not exist in the image. The extra details considered are hallucinations and thus do not contribute to necessary detail.

Figure 1. Results of the GPT-4v evaluation.

Figure 2. JSDs of the probability distributions between two adjacent layers in LLaVA-1.5.

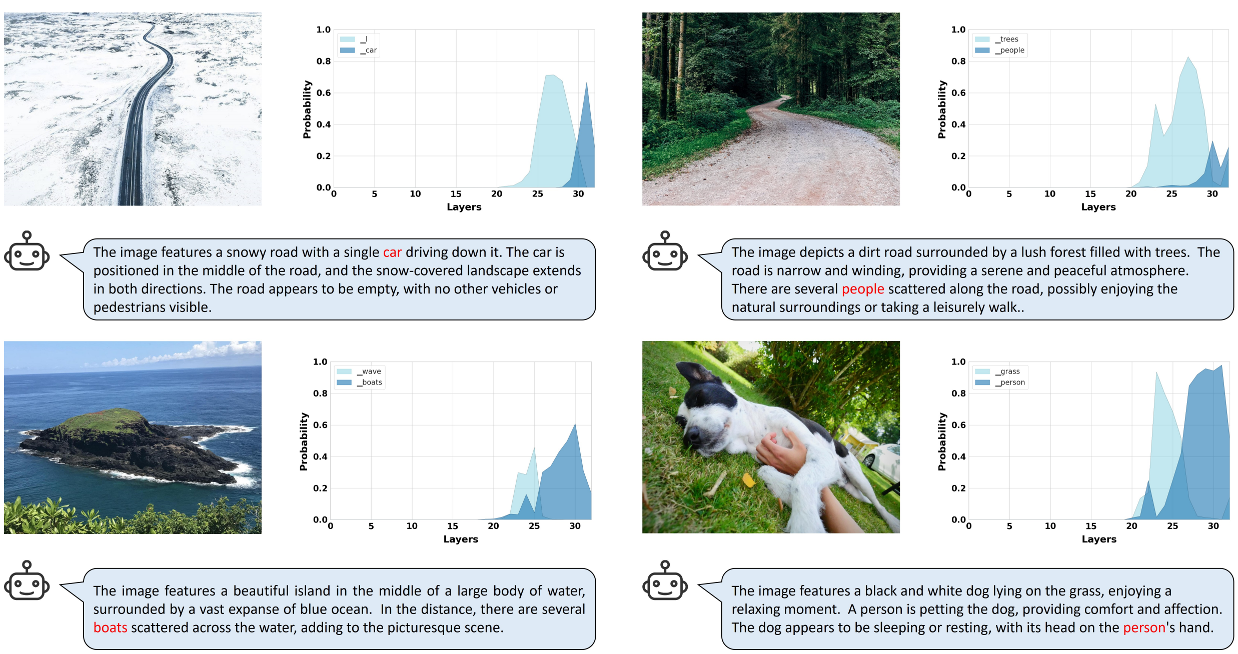


Figure 3. The probability changes of hallucinated tokens across different layers.

Please describe this image in detail.

LLaVA-1.5 (Vanilla):
The image features a man standing in a grassy field, he appears to be enjoying his time outdoors. There are several other people in the background. A car can also be seen parked in the background, slightly to the left of the center.

LLaVA-1.5 (SHIFT):
In the image, a man is standing in a grassy field, holding a colorful kite, he is standing in front of trees. Another person is visible in the background, but they are not the main focus of the scene. The field is spacious, providing plenty of room for the man to fly the kite.

Please describe this image in detail.

LLaVA-1.5 (Vanilla):
The image features two young boys playing tennis on a court. One boy is holding a tennis racket, while the other is holding a tennis ball. They are both standing on the court, ready to play. In the background, there are two cars parked, one on the left side and the other on the right side of the court.

LLaVA-1.5 (SHIFT):
The image depicts a lively scene of two young boys playing tennis on a tennis court. The two boys are standing on opposite sides of the court, each holding a tennis racket, ready to engage in a friendly match. In the background, there are two cars parked near the tennis court.

What can you see in this image?

LLaVA-1.5 (Vanilla):
The image features a woman standing next to a bicycle on a pier, overlooking a body of water, she is wearing a pink shirt and pink leggings. The bicycle is positioned close to her, with its front wheel facing the water. In the background, there are a few other people visible, but they are not the main focus of the scene.

LLaVA-1.5 (SHIFT):
The image depicts a woman standing on a pier next to a green bicycle. She is wearing tight pink exercise clothes and appears to be enjoying the view of the water. The bicycle is positioned near the edge of the pier, with the woman standing close to it.

Introduce about this image

LLaVA-1.5 (Vanilla):
The image features a bright orange sports car parked on a driveway in front of a house. The car is positioned near the center of the scene, and it appears to be the main focus of the image. In addition to the sports car, there are two people visible in the image.

LLaVA-1.5 (SHIFT):
The image features an orange sports car parked on a driveway near a house. The car is parked in front of a garage, and it appears to be parked on a gravel driveway. Trees can be seen in the background, providing a natural setting for the scene.

Figure 4. SHIFT's performance on reducing hallucinations of LLaVA-1.5.

Please describe this image in detail.

mPLUG-Owl2 (Vanilla):
The image features a man standing in a bathroom. The bathroom has a sink and a toilet, with the toilet located on the left side of the room. There are two other people in the bathroom, one standing near the left side of the room and another person closer to the center.

mPLUG-Owl2 (SHIFT):
In the image, a man is standing in a bathroom. He is standing next to a urinal, which is located on the left side of the bathroom. Behind him, another urinal can be seen on the right side. The bathroom features a brown tile wall and a brown tile floor.

Please describe this image in detail.

mPLUG-Owl2 (Vanilla):
The image features a vintage silver truck parked on a grassy field, surrounded by a mountainous landscape. The truck is positioned in the left of the scene, with a dirt road running alongside it. In the background, there are two people standing near the truck, possibly admiring its vintage design or contemplating its history.

mPLUG-Owl2 (SHIFT):
The image depicts a vintage silver truck parked on a grassy field, surrounded by a mountainous landscape. The truck is positioned on the left side of the scene, with a roadway visible in the background. The scene appears to be a rural setting, possibly a farm or ranch.

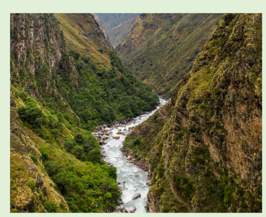
What can you see in this image?

mPLUG-Owl2 (Vanilla):
In the image, a person is holding a small piglet, which appears to be a baby pig. The piglet is being held in the person's arms, and the person is wearing a blue shirt. The scene takes place in a room, with a chair visible in the background.

mPLUG-Owl2 (SHIFT):
In the image, a person is holding a small, baby pig in their arms. The pig is cuddled up close to the person, and the person is wearing a blue shirt. The scene appears to be a close, affectionate moment between the person and the pig.

Introduce about this image

mPLUG-Owl2 (Vanilla):
The image showcases a beautiful mountain landscape with a river flowing through it. The river is surrounded by lush green trees and mountains, creating a serene and picturesque scene. There are several people scattered throughout the scene, likely enjoying the natural beauty of the area.



mPLUG-Owl2 (SHIFT):
The image depicts a breathtaking view of a river flowing through a picturesque valley. The river is surrounded by tall, green mountains, creating a serene and picturesque scene. There are several trees scattered throughout the scene, providing a sense of depth and enhancing the overall ambiance.

Figure 5. SHIFT's performance on reducing hallucinations of mPLUG-Owl2.

What is playing on the monitor?

Vanilla:
A computer mouse is playing on the monitor.

SHIFT:
There is no monitor visible in the image.

Is this photo taken in a restaurant, a kitchen, or anywhere else?

Vanilla:
The photo is taken in a restaurant, where a man is cooking food on a grill.

SHIFT:
This photo is taken in a crowded outdoor area, where a chef is cooking food on grills. It is not a restaurant or a kitchen.

What colors are the traffic signs at the top, in the middle, and at the bottom?

Vanilla:
The traffic signs at the top are blue and white, while the signs in the middle are red and white. The traffic sign at the bottom is yellow and white.

SHIFT:
The traffic signs are white, blue and white.

What is being washed in the sink?

Vanilla:
A toothbrush is being washed in the sink.

SHIFT:
Nothing is being washed in the sink. The sink is empty and clean.

Describe the person playing tennis in this photo.

Vanilla:
The person playing tennis in this photo is a woman who is holding a tennis racket and a tennis ball. She is standing on a tennis court, ready to play.

SHIFT:
There is no person visible in the image.

Figure 6. SHIFT's performance on the VQA task.