PerturboLLaVA: Reducing Multimodal Hallucination with Perturbative Visual Training

Cong Chen[1,*], Mingyu Liu[1,*], Chenchen Jing[3], Yizhou Zhou[2], Fengyun Rao[2], Hao Chen[1] , Bo Zhang[1], Chunhua Shen[3,1]

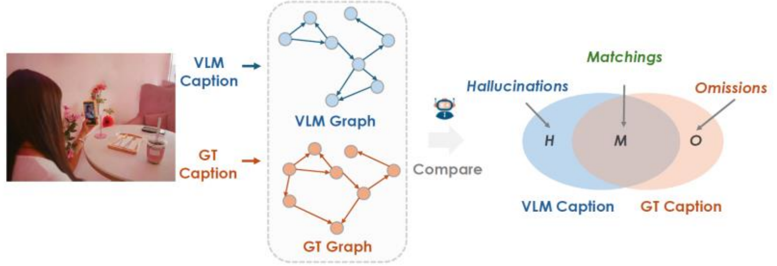
[1] Zhejiang University, China [2] WeChat Group [3] Zhejiang University of Technology

1. Introduction

- We introduce HalFscore, a novel metric built upon the language graph that is designed to evaluate both the accuracy and completeness of dense captions at a granular level.
- We identify the root cause of hallucinations in MLLMs as its inherent language bias, and propose perturbative visual training, enhancing the model's focus on visual content during training.
- The proposed method integrates seamlessly into existing training pipelines, introducing minimal additional cost. It provides a scalable, efficient solution to enhance multimodal models' visual understanding capabilities, excelling over prior compared to state-of-the-art methods across multiple dimensions.

2. Huallucination F-Score

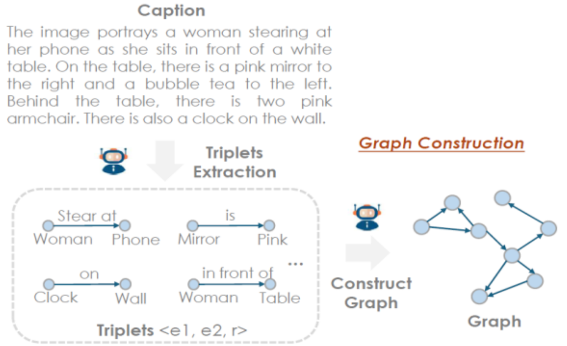
- We construct graphs for both the MLLM's output and the ground truth. By comparing the graphs, we can identify hallucinations—concepts generated by the model that contradict the ground truth, and omissions—concepts present in the ground truth but absent in the model's captions.



$$\text{Precision} = \frac{|\mathcal{C}_{\text{gen}} \cap \mathcal{C}_{\text{gt}}|}{|\mathcal{C}_{\text{gen}}|} = 1 - \frac{|\mathcal{C}_{\text{hallucinated}}|}{|\mathcal{C}_{\text{gen}}|}$$

$$\text{Recall} = \frac{|\mathcal{C}_{\text{gen}} \cap \mathcal{C}_{\text{gt}}|}{|\mathcal{C}_{\text{gt}}|} = 1 - \frac{|\mathcal{C}_{\text{omitted}}|}{|\mathcal{C}_{\text{gt}}|}.$$

$$\text{HalFscore} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}.$$


Caption

The image portrays a woman stearing at her phone as she sits in front of a white table. On the table, there is a pink mirror to the right and a bubble tea to the left. Behind the table, there is two pink armchair. There is also a clock on the wall.

- We extract triplets from the caption and build the graph accordingly. In this graph, nodes represent entity concepts e_i , and edges represent relational concepts r_{ij} between entities e_i and e_j .

3. Mitigate Hallucination via Perturbotive training

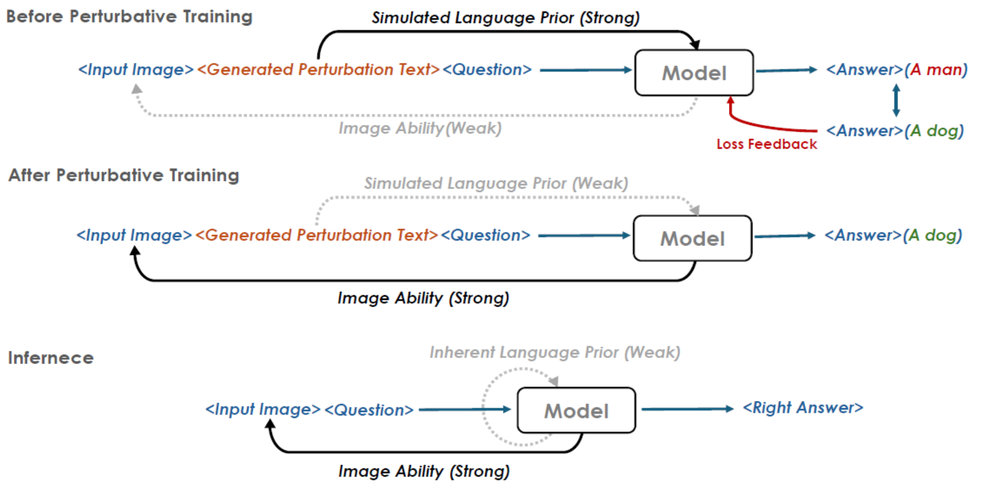
Perturbation Text Generation

<Input Image>:

<Question>: Who is sitting on the motorcycle?

Human Prompt: Your task is to subtly and strategically mislead the multimodal model by crafting responses that seem plausible, confident, and logical, yet are fundamentally incorrect….According to **<Input Image>** and **<Origin Question>.**

Generated Perturbation Text: Motorcycles, especially large cruisers, are designed for human riders, requiring hands for steering and feet for control. Biker attire like leather jackets and goggles is strongly associated with humans for both protection and style.


Before Perturbative Training
Simulated Language Prior (Strong)
<Input Image><Generated Perturbation Text><Question>
Model
<Answer>(A man)
<Answer>(A dog)
Loss Feedback
Image Ability(Weak)
After Perturbative Training
Simulated Language Prior (Weak)
<Input Image><Generated Perturbation Text><Question>
Model
<Answer>(A dog)
Image Ability (Strong)
Infernece
Inherent Language Prior (Weak)
<Input Image><Question>
Model
<Right Answer>
Image Ability (Strong)

- **Method overview:** we perturbations in the textual inputs during training. This approach simulates the effect of language priors and forces the model to adjust its responses based on visual data rather than textual biases.
- **Perturbation Text Design:** 1) Contextual relevance. Perturbation is expected to be contextually relevant to the image content 2) Alignment with pretrained knowledge. 3) Semantic variation.
- **Detailed training:** we insert perturbation text to sft data, after system-prompts and before QA to substitute sft training

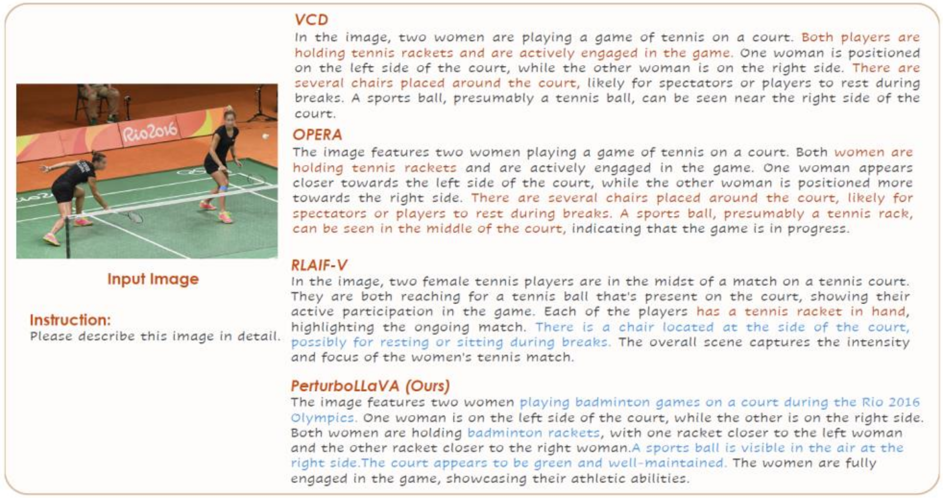
4. Main Results

- **Qualitative Results of metric evaluation**

Model	Size	Precision↑	Recall↑	Fscore↑	Object↓	Attribute↓	Relation↓
Ovis1.6-Gemma2	9B	61.5	50.3	55.4	22.1	4.9	11.7
Qwen2-VL	7B	60.8	50.0	54.9	24.2	5.1	9.9
LLaVa-onevision	7B	61.3	48.3	54.1	22.1	4.9	11.7
InternVL2	8B	60.6	48.6	53.9	24.1	5.2	10.1
Idefics3	8B	59.7	48.2	53.3	25.1	6.1	9.1
MiniCPM-2.6	8B	57.3	47.8	52.1	26.3	6.7	9.7
LLaVA1.5	7B	53.3	45.8	49.2	28.1	8.1	10.5
RLAIF-V*	7B	57.7	47.2	51.9	25.9	7.0	9.4
OPERA*	7B	58.6	46.7	51.9	25.3	6.9	9.2
VCD*	7B	58.8	46.7	52.0	25.6	7.2	8.4
Ours*	7B	59.5	46.5	52.2	25.3	6.4	8.8

- **Qualitative Results of perturbotive training**

Model	Size	Reward	HalFscore			Object HalBench		HalBench↑	MMB↑	SEED↑	CCBench↑
			Prec.↑	Recall↑	Fscore↑	CHAIR$_s$↓	CHAIR$_i$↓				
LLaVA1.5	7B	✗	53.3	45.8	49.2	54.2	15.0	46.9	67.3	65.3	29.4
VCD	7B	✗	58.8	46.7	52.0	51.2	14.3	46.9	66.2	64.7	29.4
OPERA	7B	✗	58.6	46.7	51.9	50.8	14.2	47.1	67.4	65.3	29.4
RLAI-F	7B	LLaVA-Next	57.7	**47.2**	51.9	**18.1**	**4.7**	**51.3**	63.7	65.4	27.8
Ours	7B	✗	59.5	46.5	52.2	36.1	10.4	47.5	68.9	65.6	30.6
OPERA+Ours	7B	✗	**60.2**	47.0	**52.8**	33.1	10.1	47.6	**68.9**	**65.6**	**31.0**



Instruction:
Please describe this image in detail.

VCD
In the image, two women are playing a game of tennis on a court. Both players are holding tennis rackets and are actively engaged in the game. One woman is positioned on the left side of the court, while the other woman is on the right side. There are several chairs placed around the court, likely for spectators or players to rest during breaks. A sports ball, presumably a tennis ball, can be seen near the right side of the court.

OPERA
The image features two women playing a game of tennis on a court. Both women are holding tennis rackets and are actively engaged in the game. One woman appears closer towards the left side of the court, while the other woman is positioned more towards the right side. There are several chairs placed around the court, likely for spectators or players to rest during breaks. A sports ball, presumably a tennis rack, can be seen in the middle of the court, indicating that the game is in progress.

RLAIF-V
In the image, two female tennis players are in the midst of a match on a tennis court. They are both reaching for a tennis ball that's present on the court, showing their active participation in the game. Each of the players has a tennis racket in hand, highlighting the ongoing match. There is a chair located at the side of the court, possibly for resting or sitting during breaks. The overall scene captures the intensity and focus of the women's tennis match.

PerturboLLaVA (Ours)
The image features two women playing badminton games on a court during the Rio 2016 Olympics. One woman is on the left side of the court, while the other is on the right side. Both women are holding badminton rackets, with one racket closer to the left woman and the other racket closer to the right woman.A sports ball is visible in the air at the right side.The court appears to be green and well-maintained. The women are fully engaged in the game, showcasing their athletic abilities.