From Multimodal LLM to Human-level AI

***Modality*, *Instruction*, *Reasoning*, *Efficiency* and Beyond**

https://mllm2024.github.io/CVPR2024/

Hao Fei

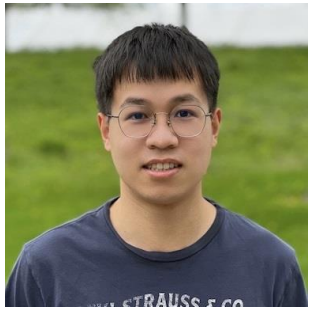
National University of Singapore

Yuan Yao

National University of Singapore

Ao Zhang

National University of Singapore

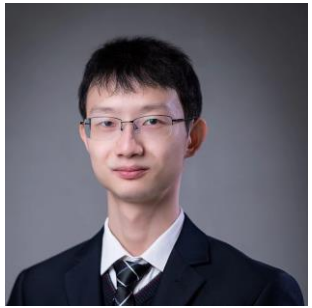


Haotian Liu

University of Wisconsin-Madison

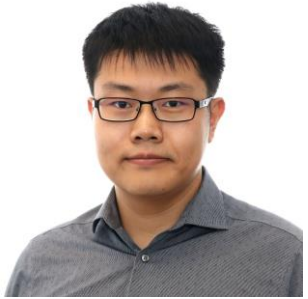
Fuxiao Liu

University of Maryland, College Park



Zhuosheng Zhang

Shanghai Jiao Tong University



Hanwang Zhang

Nanyang Technological University

Shuicheng Yan

Kunlun 2050 Research, Skywork AI

Part-V

Multimodal Hallucinations

Fuxiao Liu

Ph.D.

University of Maryland, College Park

https://fuxiaoliu.github.io

Table of Content

1

What’s hallucination in MLLMs ?

Definition of Hallucinations in MLLMs

Hallucination of MLLM generally refers to the phenomenon where the *generated text response* *does not align with* the *visual content*.

Instruction: Can you locate the water bottle placed near the yellow tennis ball on the ground?

MiniGPT4-v1: *The water bottle is located near the yellow tennis ball on the ground. Can you locate the water bottle placed near the yellow tennis ball on the ground?*

MiniGPT4-v2: *Yes, the water bottle is placed near the yellow tennis ball on the ground.*

LLaVA1.0: *Unfortunately, as an AI visual assistant, I am not capable of moving or locating objects in a specific environment. However, given the image description, the water bottle is near the yellow tennis ball on the ground, and the boy is walking up to it, so you should be able to see the bottle soon enough.*

LLaVA1.5: *Yes, there is a water bottle placed near the yellow tennis ball on the ground.*

[1] Mitigating Hallucination in Large Multi-Modal Models via Robust Instruction Tuning. ICLR 2024

Definition of Hallucinations in MLLMs

Hallucination of MLLM generally refers to the phenomenon where the *generated text response* *does not align with* the *visual content*.

Object Hallucination:

There are some benches and a fence in the background.

Attribution Hallucination:

There is a large tree in the background with pink flowers.

Relation Hallucination:

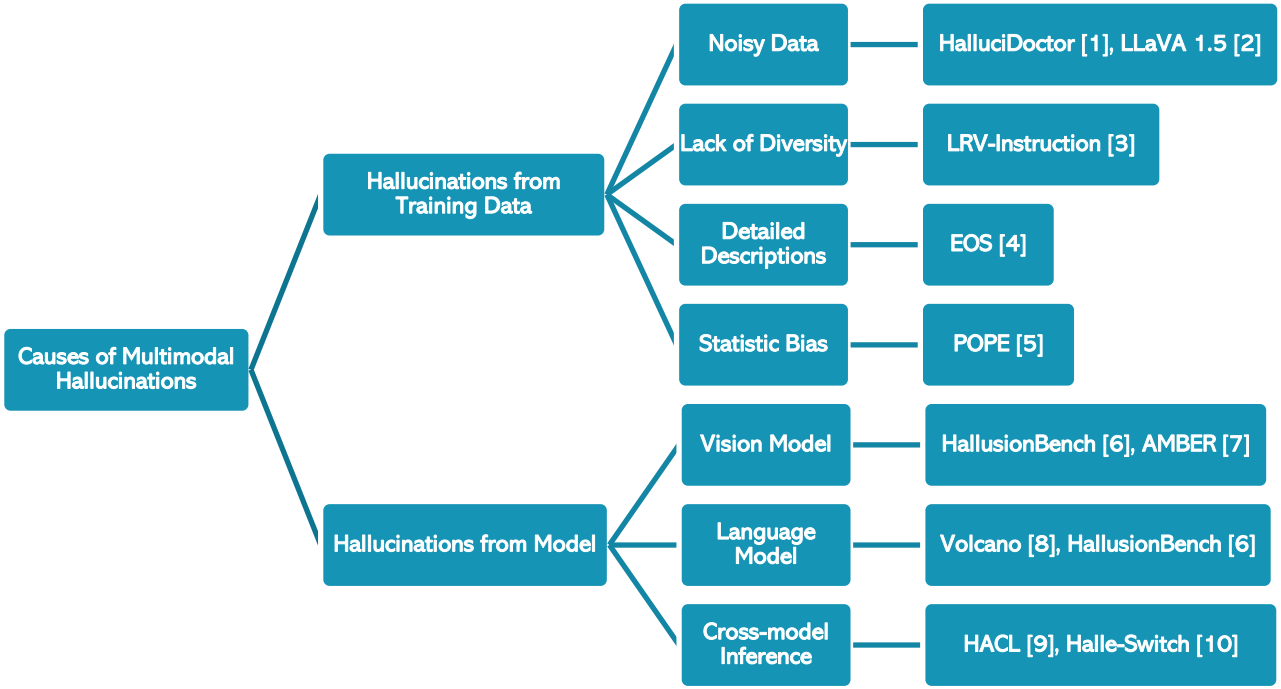
The other people in the picture are standing around the girl, watching what she is doing.

[1] Hallucination of Multimodal Large Language Models: A Survey. 2024

2

Causes of Hallucinations in MLLMs

Causes of Hallucinations in MLLMs



[1] HalluciDoctor: Mitigating Hallucinatory Toxicity in Visual Instruction Data. CVPR 2024.
[2] Improved baselines with visual instruction tuning. CVPR 2024..
[3] Mitigating hallucination in large multi-modal models via robust instruction tuning. ICLR 2024..
[4] Mitigating Multimodal Hallucination from an EOS Decision Perspective. ACL 2024.
[5] Evaluating object hallucination in large vision-language models. EMNLP 2023..
[6] HallusionBench: You See What You Think? Or You Think What You See? An Image-Context Reasoning Benchmark Challenging for GPT-4V(Ision), LLaVA-1.5, and Other Multi-Modality Models. CVPR 2024.
[7] An llm-free multi-dimensional benchmark for mllms hallucination evaluation. 2023.
[8] Volcano: Mitigating Multimodal Hallucination through Self-Feedback Guided Revision. 2023.
[9] Hallucination Augmented Contrastive Learning for Multimodal Large Language Model. 2023.
[10] HallE-Switch: Controlling Object Hallucination in Large Vision Language Models. 2023.

Noisy Data

- Pre-training stage employs image-text pairs crawled from the web, which contain inaccurate information.
- As for instruction tuning data, LLaVA utilizes the advanced text-only GPT-4 model to generate instructions. However, text-only ChatGPT is a language model that cannot interpret visual content, leading to the risk of noisy data.

LLaVA: At a train station, a group of people, including both young children and adults, are standing on a platform waiting for a train to arrive. The train is already present on the tracks, partially visible on the right side of the image. Some of the people watch the train closely, while others seem to be patiently anticipating its departure.

There is a total of eight individuals waiting for the train, with one child in the middle of the platform and the others scattered around. A backpack can be found on the far left side of the platform, suggesting that someone may have set it down while waiting.

[1] ShareGPT4V: Improving Large Multi-Modal Models with Better Captions. 2023.

Lack of Data Diversity

- Most instruction following data samples are composed of positive instruction, without negative instruction data and reject answering responses.

Results on POPE

Model	Yes(%)
mPLUG-Owl	96.23
LLaVA	95.37
MultiModel-GPT	99.97

Tend to answer "Yes" !

(Neg) Nonexistent Object Manipulation

Q: Can you spot a hot air balloon above the tall tree?
A: There is no hot air balloon mentioned in the image.

(Neg) Existent Object Manipulation

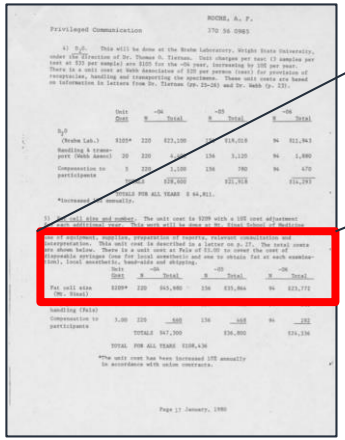
Q: Is the woman in the blue pants and pink shirt sitting on the floor?
A: There is no hot air balloon mentioned in the image.

Examples from LRV-Instruction

[1] Mitigating hallucination in large multi-modal models via robust instruction tuning. ICLR 2024.
[2] Evaluating Object Hallucination in Large Vision-Language Models. EMNLP 2023.

Hallucinations from Vision Model

- A weak vision model can lead to misinterpretation of visual content.

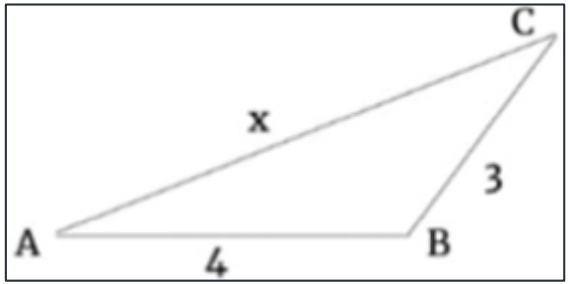


are shown below. There is a unit cost at Fels of $3.00 to cover the cost of disposable syringes (one for local anesthetic and one to obtain fat at each examination), local anesthetic, band-aids and shipping.

	Unit Cost	-04 N	-04 Total	-05 N	-05 Total	-06 N	-06 Total
Fat cell size (Mt. Sinai)	$209*	220	$45,980	156	$35,864	94	$23,772

Question: What is the total cost for Fat cell size (Mt. Shlnai) in the -05 year?

Prediction: $12,800



Question: Is the value of x in the image 5?

Prediction: To determine the value of x in the triangle, we can use the Pythagorean theorem for right triangles, Where a and b are the lengths of the two shorter sides (legs) and is the length of the hypotenuse. So, yes, the value of x in the image is 5.

[1] HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. CVPR 2024.

Hallucinations from Language Model

- The language model is much larger and stronger than the vision model, leading to a tendency to prioritize its parametric memory.

	Gold	Silver	Bronze
China	23	21	28
United States	36	38	36
Russian Fed.	23	21	28
Great Britain	19	13	15
Germany	16	10	15
Australia	14	15	17
Korea	13	10	8
Japan	9	6	10
Italy	8	10	10
France	7	16	17

The gold medal number of China is changed to 23, ranked second.

Question: According to the metal number chart of the 2008 Beijing Olympics in the image, Does China have the most gold metals, followed by the USA and Russia?

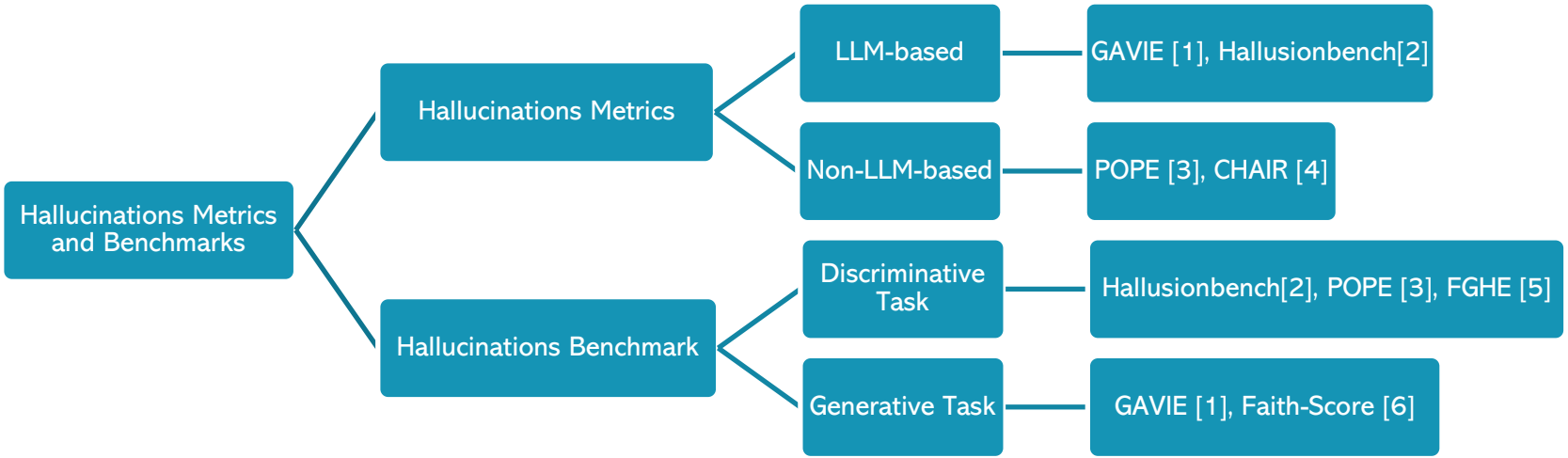
Prediction: Yes, according to the metal number chart of the 2008 Beijing Olympic Games, China won the most gold medals, followed by the United States and Russia. The chart shows the number of gold, silver, and bronze medals won by each country during the games.

[1] HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. CVPR 2024.

3

Multimodal Hallucination Metrics and Benchmarks

Multimodal Hallucination Metric & Benchmarks



[1] Mitigating hallucination in large multi-modal models via robust instruction tuning. ICLR 2024.
[2] HallusionBench: You See What You Think? Or You Think What You See? An Image-Context Reasoning Benchmark Challenging for GPT-4V(Ision), LLaVA-1.5, and Other Multi-Modality Models. CVPR 2024.
[3] Evaluating object hallucination in large vision-language models. EMNLP 2023.
[4] Object hallucination in image captioning. EMNLP 2018.
[5] Mitigating Fine-Grained Hallucination by Fine-Tuning Large Vision-Language Models with Caption Rewrites. MMM 2024.
[6] FAITHSCORE: Evaluating Hallucinations in Large Vision-Language Models. 2023

Multimodal Hallucination Metric & Benchmarks

Benchmark	Venue	Underlying Data Source	Size	Task Type	Metric	Hallucination Type			
						Category	Attribute	Relation	Others
CHAIR [90]	EMNLP'18	MSCOCO [70]	5,000	Gen	CHAIR	✓	✗	✗	✗
POPE [69]	EMNLP'23	MSCOCO [70]	3,000	Dis	Acc/P/R/F1	✓	✗	✗	✗
MME [113]	arXiv'23 Jun	MSCOCO [70]	1457	Dis	Acc/Score	✓	✓	✗	✓
CIEM [42]	NeurIPS-W'23	MSCOCO [70]	78120	Dis	Acc	✓	✗	✗	✗
M-HalDetect [32]	arXiv'23 Aug.	MSCOCO [70]	4,000	Dis	Reward Model Score	✓	✗	✗	✗
MMHal-Bench [96]	arXiv'23 Sep.	Open-Images [61]	96	Gen	LLM Assessment	✓	✗	✗	✓
GAVIE [73]	ICLR'24	Visual-Genome [59]	1,000	Gen	LLM Assessment	Not Explicitly Stated			
NOPE [77]	arXiv'23 Oct.	Open-Images [61]	36,000	Dis	Acc/METEOR [3]	✓	✗	✗	✗
HaELM [104]	arXiv'23 Oct.	MSCOCO [70]	5,000	Gen	LLM Assessment	Not Explicitly Stated			
FaithScore [55]	arXiv'23 Nov.	MSCOCO [70]	2,000	Gen	FaithScore	✓	✓	✓	Obj. Counting
Bingo [21]	arXiv'23 Nov.	Unknown	370	Gen	Human Assessment	✗	✗	✗	Model Bias
AMBER [103]	arXiv'23 Nov.	Web	15,202	Dis & Gen	AMBER Score	✓	✓	✓	✗
RAH-Bench [16]	arXiv'23 Nov.	MSCOCO [70]	3,000	Dis	False Positive Rate	✓	✓	✓	✗
HallusionBench [72]	CVPR'24	Unknown	1,129	Gen	LLM Assessment	✗	✗	✗	Model Diagnose
CCEval [123]	arXiv'23 Dec.	Visual-Genome [59]	100	Gen	LLM-based CHAIR	✓	✗	✗	✗
MERLIM [100]	arXiv'23 Dec.	MSCOCO [70]	31,373	Dis	Accuracy	✓	✗	✓	Obj. Counting
FGHE [105]	arXiv'23 Dec.	MSCOCO [70]	200	Dis	Acc/P/R/F	✓	✓	✓	Obj. Behavior
MOCHa [5]	arXiv'23 Dec.	Synthetic	2,000	Gen	OpenCHAIR [5]	✓	✓	✗	✗
CorrelationQA [35]	arXiv'24 Feb.	Synthetic	7,308	Dis	Acc/AccDrop	✗	✗	✗	Model Bias
VQAv2-IDK [11]	arXiv'24 Feb.	VQAv2 [30]	6,624	Dis	Acc	✗	✗	✗	IK [11]
MHaluBench [13]	arXiv'24 Feb.	MSCOCO [70]	1,860	Gen	Acc/P/R/F	✓	✓	✗	T2I
VHTest [46]	arXiv'24 Feb.	MSCOCO [70]	1,200	Dis & Gen	Acc	✓	✓	✗	✓
Hal-Eavl [53]	arXiv'24 Feb.	MSCOCO [70] & LAION [92]	10,000	Dis & Gen	Acc/P/R/F & LLM Assessment	✓	✓	✓	Obj. Event

Dis* *means: converting the evaluation of hallucination into a* ***<u>binary classification task by prompting MLLMs with simple Yes-or-No</u>*** *short questions about the probing objects . It's not open-ended questions.*

Gen* *means generative tasks, which is open-ended questions.*

[1] Hallucination of Multimodal Large Language Models: A Survey. 2024

Multimodal Hallucination Metric & Benchmarks

Non-LLM evaluation, Discriminative Tasks

Q: Is there a bottle in the image?

A: Yes, there is a bottle in the image.

Q: Is there a Knife in the image?

A: No, there is no knife in the image.

LLM evaluation, Generative Tasks

Can you help me find a dog playing with a ball on the shore?

There is a dog playing with a ball on the shore. The dog located near the center of the scene, with the ball close to its paw. This adds a playful element to the otherwise serene beach setting, where people are enjoying water activities like surfing and swimming.

mPLUG-Owl

The image is a photo of a group of three people surfing on the ocean. They are wearing black wet suits and have their boards on their heads. The sky in the background is a light blue color with white clouds.

MiniGPT4

Evaluate

Evaluate

GAVIE	
Relevancy Score(0-10)	**Accuracy Score(0-10)**
Score *(mPLUG-Owl)*: 10 **Reason**: *It followed the instruction and directly described a dog playing with a ball on the shore, which was the exact scenario asked for in the instruction.*	**Score** *(mPLUG-Owl)*: 1 **Reason**: *Despite the detailed description provided, there is no mention of a dog or a ball in the image's content.*
Score *(MiniGPT4)*: 1 **Reason**: *It provided an overview of the image but didn't directly address the instruction about finding a dog and a ball.*	**Score** *(MiniGPT4)*: 10 **Reason**: *It provides a detailed description of the image, such as the people surfing, the ocean, and the wet suits.*
Score *(Ours)*: 10 **Reason**: *The answer directly responds to the question.*	**Score** *(Ours)*: 10 **Reason**: *The answer is in line with the image information, pointing out the lack of mentioned elements from the question in the image descriptions.*

Finetune

LRV-Instruction Dataset(Ours)

There is no mention of a dog or a ball in the given image. Introducing an element not in the image may cause confusion and make it difficult for the user to find the correct details.

[1] Mitigating hallucination in large multi-modal models via robust instruction tuning. ICLR 2024.
[3] Evaluating object hallucination in large vision-language models. EMNLP 2023.

Multimodal Hallucination Metric & Benchmarks

Benchmark	Venue	Underlying Data Source	Size	Task Type	Metric	Hallucination Type			
						Category	Attribute	Relation	Others
CHAIR [90]	EMNLP'18	MSCOCO [70]	5,000	Gen	CHAIR	✓	✗	✗	✗
POPE [69]	EMNLP'23	MSCOCO [70]	3,000	Dis	Acc/P/R/F1	✓	✗	✗	✗
MME [113]	arXiv'23 Jun	MSCOCO [70]	1457	Dis	Acc/Score	✓	✓	✗	✓
CIEM [42]	NeurIPS-W'23	MSCOCO [70]	78120	Dis	Acc	✓	✗	✗	✗
M-HalDetect [32]	arXiv'23 Aug.	MSCOCO [70]	4,000	Dis	Reward Model Score	✓	✗	✗	✗
MMHal-Bench [96]	arXiv'23 Sep.	Open-Images [61]	96	Gen	LLM Assessment	✓	✗	✗	✓
GAVIE [73]	ICLR'24	Visual-Genome [59]	1,000	Gen	LLM Assessment	Not Explicitly Stated			
NOPE [77]	arXiv'23 Oct.	Open-Images [61]	36,000	Dis	Acc/METEOR [3]	✓	✗	✗	✗
HaELM [104]	arXiv'23 Oct.	MSCOCO [70]	5,000	Gen	LLM Assessment	Not Explicitly Stated			
FaithScore [55]	arXiv'23 Nov.	MSCOCO [70]	2,000	Gen	FaithScore	✓	✓	✓	Obj. Counting
Bingo [21]	arXiv'23 Nov.	Unknown	370	Gen	Human Assessment	✗	✗	✗	Model Bias
AMBER [103]	arXiv'23 Nov.	Web	15,202	Dis & Gen	AMBER Score	✓	✓	✓	✗
RAH-Bench [16]	arXiv'23 Nov.	MSCOCO [70]	3,000	Dis	False Positive Rate	✓	✓	✓	✗
HallusionBench [72]	CVPR'24	Unknown	1,129	Gen	LLM Assessment	✗	✗	✗	Model Diagnose
CCEval [123]	arXiv'23 Dec.	Visual-Genome [59]	100	Gen	LLM-based CHAIR	✓	✗	✗	✗
MERLIM [100]	arXiv'23 Dec.	MSCOCO [70]	31,373	Dis	Accuracy	✓	✗	✓	Obj. Counting
FGHE [105]	arXiv'23 Dec.	MSCOCO [70]	200	Dis	Acc/P/R/F	✓	✓	✓	Obj. Behavior
MOCHa [5]	arXiv'23 Dec.	Synthetic	2,000	Gen	OpenCHAIR [5]	✓	✓	✗	✗
CorrelationQA [35]	arXiv'24 Feb.	Synthetic	7,308	Dis	Acc/AccDrop	✗	✗	✗	Model Bias
VQAv2-IDK [11]	arXiv'24 Feb.	VQAv2 [30]	6,624	Dis	Acc	✗	✗	✗	IK [11]
MHaluBench [13]	arXiv'24 Feb.	MSCOCO [70]	1,860	Gen	Acc/P/R/F	✓	✓	✗	T2I
VHTest [46]	arXiv'24 Feb.	MSCOCO [70]	1,200	Dis & Gen	Acc	✓	✓	✗	✓
Hal-Eavl [53]	arXiv'24 Feb.	MSCOCO [70] & LAION [92]	10,000	Dis & Gen	Acc/P/R/F & LLM Assessment	✓	✓	✓	Obj. Event

Manually crafted

Many benchmarks are sourced from MSCOCO and visual-Genome, which are usually included in current instruction tuning datasets.
***HallusionBench** manually crafted by human experts, is an ideal benchmark for the zero-shot evaluation.*

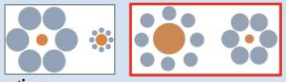
[1] Hallucination of Multimodal Large Language Models: A Survey. 2024

Multimodal Hallucination Metric & Benchmarks

HallusionBench: *manually crafted by human experts, is an ideal benchmark for the zero-shot evaluation* ***with more diverse image types, hallucination types.***

Visual Dependent

Illusion



Question:
Is the right orange circle the same size as the left orange circle?
Is the right orange circle larger than the left orange circle?
Is the right orange circle smaller than the left orange circle?

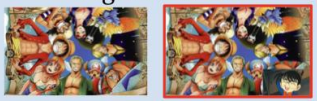
Math

Question:
According to parallel lines theorem, is angle 1 + angle 2 > 180 ?
According to parallel lines theorem, is angle 1 + angle 2 = 180 ?
According to parallel lines theorem, is angle 1 + angle 2 < 180 ?

Poster

Question:
Does the image show "Beijing Roast Duck"?
Does the image show "Guangxi Roast Duck"?

Figure / Other



Question:
Are all the characters in this figure from the manga series One Piece?
Are there any characters in this figure from the manga series Detective Conan?

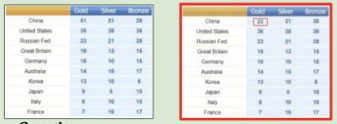
Video

Question:
According to the positive sequence images, does Homer Simpson disappear into the bushes?
According to the positive sequence images, does Homer Simpson come out of the bushes?
Homer Simpson disappears into the bushes. According to the positive sequence, are they in the correct order?
Homer Simpson comes out of the bushes. According to the positive sequence, are they in the correct order?

Visual Supplement

Table

No Visual



	Gold	Silver	Bronze
China	51	21	28
United States	36	38	36
Russian Fed.	23	21	28
Great Britain	19	13	15
Germany	16	10	15
Australia	14	15	17
Korea	13	10	8
Japan	9	6	10
Italy	8	10	10
France	7	16	17

	Gold	Silver	Bronze
China	36	21	28
United States	36	38	36
Russian Fed.	23	21	28
Great Britain	19	13	15
Germany	16	10	15
Australia	14	15	17
Korea	13	10	8
Japan	9	6	10
Italy	8	10	10
France	7	16	17

Question:
Does China have the most gold medals in 2008 beijing olympic?
Does USA have the most gold medals in 2008 beijing olympic?
Does Russia have the most gold medals in 2008 beijing olympic?

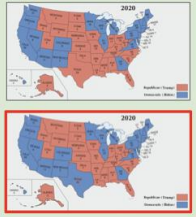
Chart

No Visual

Question:
In 2017, was Tencent the company with the highest revenue from video games, with Sony as the second-highest earner?
In 2017, did Apple generate higher revenue from video games compared to Google?

Map

No Visual



Question:
Based on the map, did the Democratic Party win Texas in the 2020 elections?
Based on the map, did the Republican Party win Texas in the 2020 elections?

OCR

No Visual

$G \approx 6.67428 \times 10^{-11}\, m^3\, kg^{-1}\, s^{-2}$

$G \approx 6.6\ 9428 \times 10^{-11}\, m^3\, kg^{-1}\, s^{-2}$

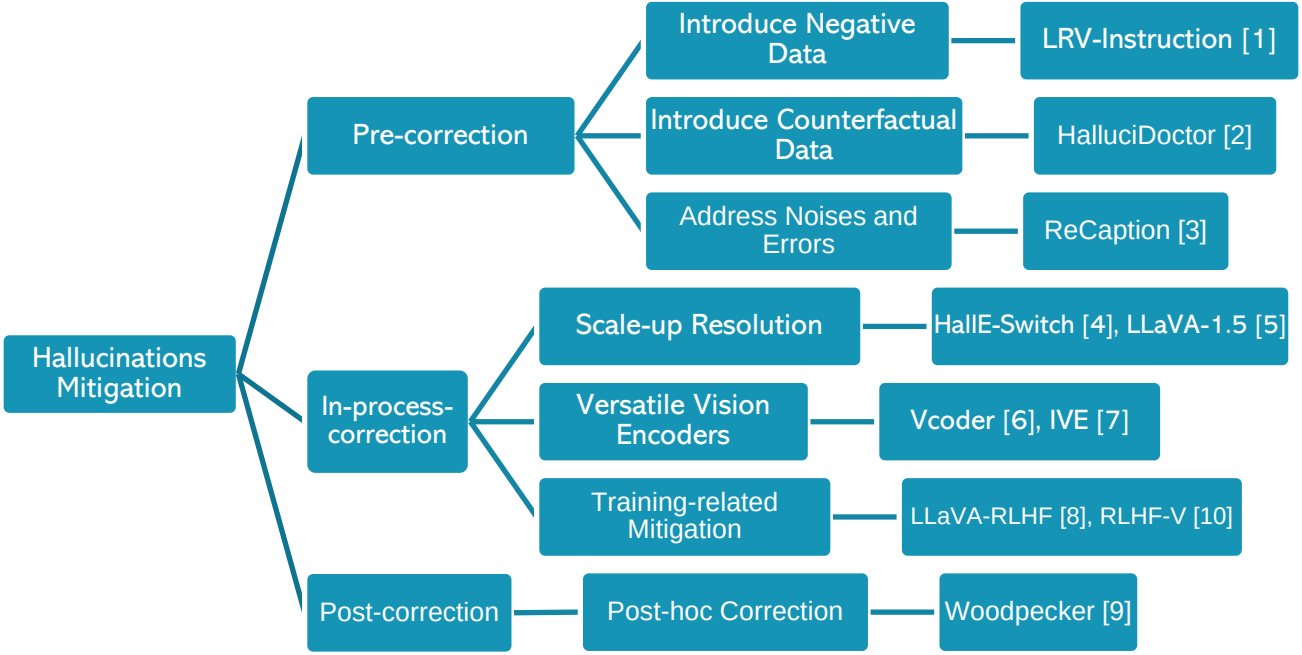
Question:
*According to the image, does the value of Gravity constant 'G' range from 6.66 * 10^-11 to 6.68 * 10^-11?*
*According to the image, does the value of Gravity constant 'G' range from 6.68 * 10^-11 to 6.70 * 10^-11?*

[1] HallusionBench: An Advanced Diagnostic Suite for Entangled Language Hallucination and Visual Illusion in Large Vision-Language Models. CVPR 2024.

4

Multimodal Hallucination Mitigation

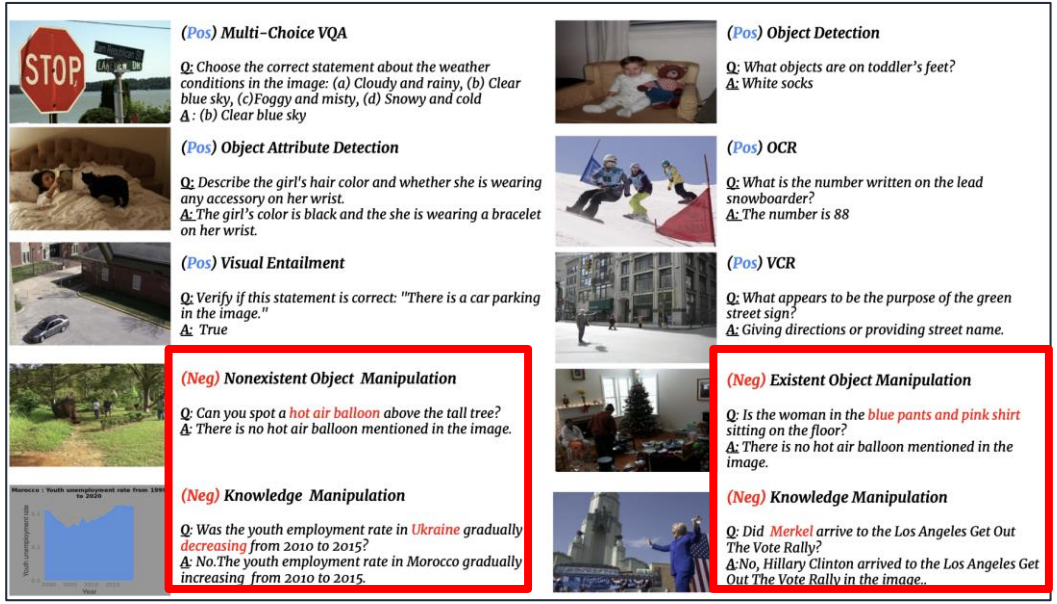
Multimodal Hallucination Mitigation



[1] Mitigating Multimodal Hallucination from an EOS Decision Perspective. ACL 2024..
[2] HalluciDoctor: Mitigating Hallucinatory Toxicity in Visual Instruction Data. 2023.
[3] Mitigating Fine-Grained Hallucination by Fine-Tuning Large Vision-Language Models with Caption Rewrites. MMM 2024..
[4] HallE-Switch: Controlling Object Hallucination in Large Vision Language Models. 2023.
[4] Object hallucination in image captioning. 2023
[5] Improved baselines with visual instruction tuning. CVPR 2024..
[6] Vcoder: Versatile vision encoders for multimodal large language models. 2023.
[7] Incorporating Visual Experts to Resolve the Information Loss in Multimodal Large Language Models. 2024.
[8] Aligning large multimodal models with factually augmented rlhf. 2023.
[9] Woodpecker: Hallucination Correction for Multimodal Large Language Models. 2023.
[10] RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback. CVPR 2024.

Introduce Negative Data

- ***LRV-Instruction*** is designed to include both positive and negative instructions for more robust visual instruction tuning.



Results on MME Benchmark

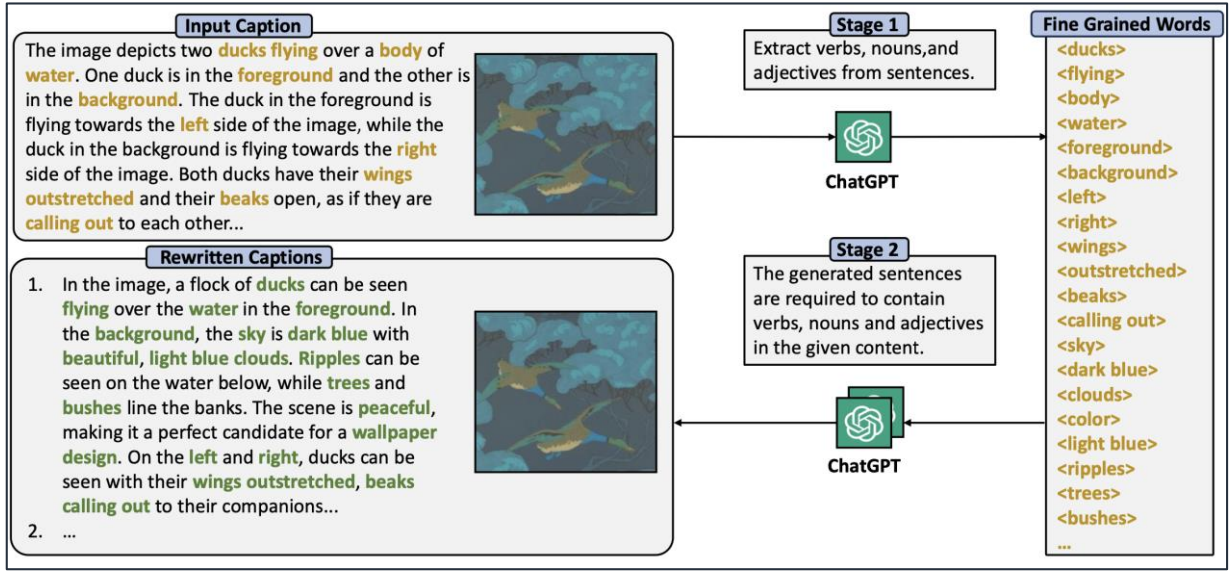
Backbone	Perception	Cognition
Original MiniGPT4	616.41	232.71
Finetuned MiniGPT4	**895.96**	**296.43**
Original mPLUG-Owl	967.34	276.07
Finetuned mPLUG-Owl	**1298.78**	**328.21**

Backbone	Acc(Pos)	Acc(Neg)
Original MiniGPT4	0.53	0.54
Finetuned MiniGPT4	**0.58**	**0.68**
Original mPLUG-Owl	0.62	0.55
Finetuned mPLUG-Owl	**0.69**	**0.78**

[1] Mitigating hallucination in large multi-modal models via robust instruction tuning. ICLR 2024..
[2] MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models.

Address Noises and Errors

- Rewrite the text captions of existing image-text pairs in datasets by ChatGPT.



Evaluation over different hallucination categories in terms of F1 score of FGHE.

Method	Multi-Object	Attribute	Behavior
mPLUG-Owl	72.36	68.14	61.59
w/ ReCaption	**74.23**	**71.55**	**67.90**
	(+1.87)	(+3.41)	(+6.31)
LLaVA	74.26	71.49	66.45
w/ ReCaption	**76.70**	**75.18**	**71.82**
	(+2.44)	(+3.69)	(+5.37)
MultiModal-GPT	74.82	70.56	68.13
w/ ReCaption	**76.84**	**78.92**	**74.22**
	(+2.02)	(+8.36)	(+6.09)
MiniGPT-4	63.30	60.16	56.72
w/ ReCaption	**71.02**	**72.16**	**67.93**
	(+7.72)	(+12.00)	(+11.21)

[1] Mitigating Fine-Grained Hallucination by Fine-Tuning Large Vision-Language Models with Caption Rewrites. MMM 2024..

Scale-up Resolution

- Higher resolution generally results in lower degrees of hallucination.

Experiment results on CCEval [1] of LLaVA with Llama2 13B

Benchmark	Vision Encoder	$CHAIR_s \downarrow$	$CHAIR_i \downarrow$	Coverage↑	Avg. Length↑	Avg. Object↑
CCEval (Ours)	CLIP-L-112x	79.00	21.70	32.04	110.36	9.12
	CLIP-L-224x	74.00	19.30	32.83	113.03	9.18
	CLIP-L-336x	64.00	16.00	33.37	108.52	8.92

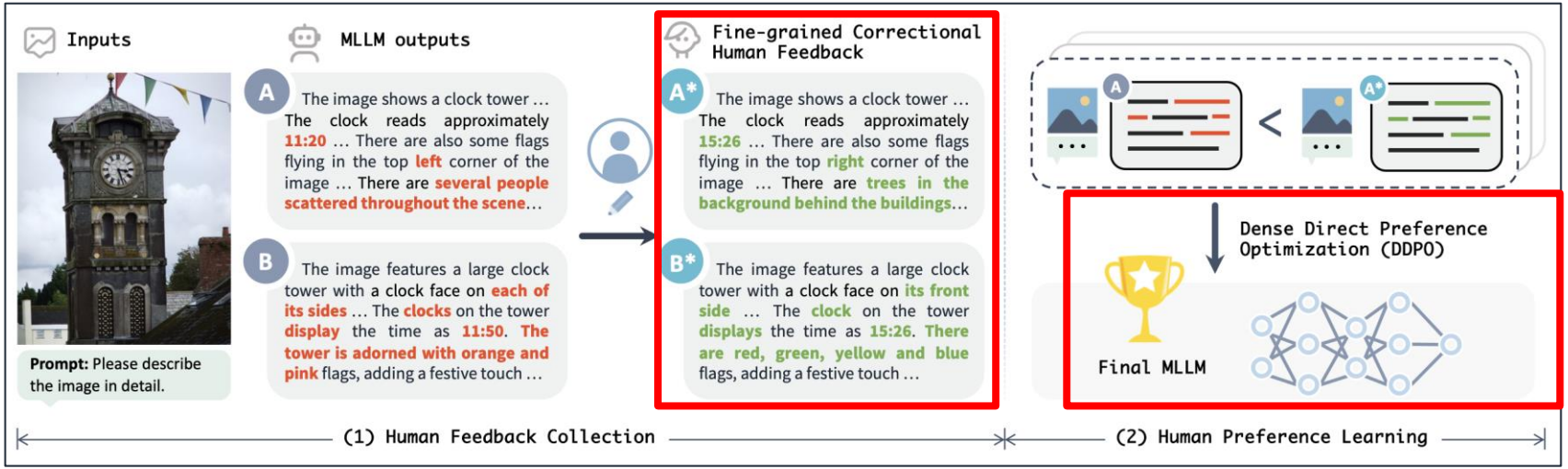
Experiment results on VQAv2, GQA, VisWiz and TextVQA of LLaVA 1.5

	LLM	Resolution	VQAv2	GQA	VisWiz	TextVQA
LLaVA 1.5	Vicuna-13B	336	80.0	63.3	53.6	61.3
LLaVA 1.5	Vicuna-13B	448	81.8	64.7	57.5	62.5

[1] HALLE-SWITCH: RETHINKING AND CONTROLLING OBJECT EXISTENCE HALLUCINATIONS IN LARGE VISION-LANGUAGE MODELS FOR DETAILED CAPTION. 2023.
[2] Improved baselines with visual instruction tuning. CVPR 2024.

Training-related Mitigation: RLHF-V

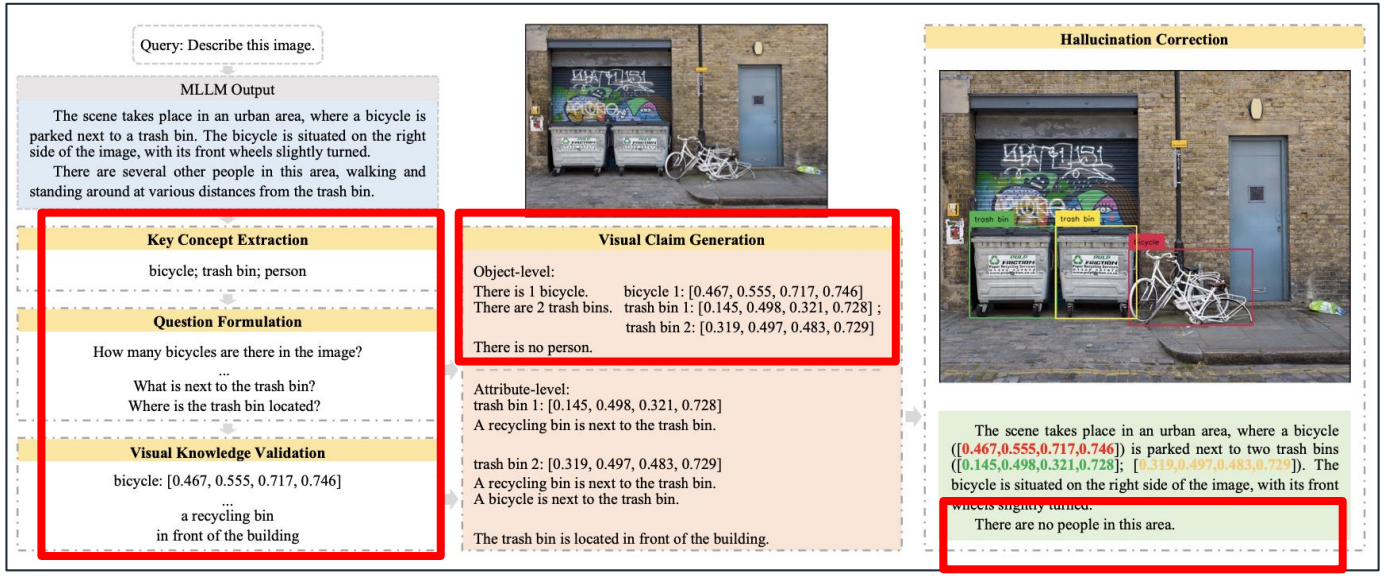
- Collect human feedback in the form of fine-grained segment-level corrections.
- Propose DDPO to optimizes the policy model against dense and fine-grained segment-level preference.



[1] RLHF-V: Towards Trustworthy MLLMs via Behavior Alignment from Fine-grained Correctional Human Feedback. CVPR 2024..

Post-hoc Correction

- ***Woodpecker*** is training-free general framework for hallucination correction. It incorporates expert models to supplement contextual information of the image and crafts a pipeline to correct hallucinations step by step.



Experiment results on MME

Model	w/ correction	Total
LLaVA	No	421
LLaVA	Yes	**565**
Otter	No	448
Otter	Yes	**571**

[1] Woodpecker: Hallucination Correction for Multimodal Large Language Models. 2023.
[2] MME: A Comprehensive Evaluation Benchmark for Multimodal Large Language Models. 2023.

Future Directions of MLLM Hallucinations

Establishing Standardized Benchmarks

- *Easy to use, fair, free-form, cheap*

Reframing Hallucination as a Feature

- *It's only when the dreams enter deemed factually incorrect territory that we label them as 'hallucinations'.*
- *Double-edged sword / Creation / Hallucination*

Any questions?

You can find me at:

- fl3es@umd.edu