

### From Multimodal LLM to Human-level Al *Evaluations and Benchmarks*





https://mllm2024.github.io/CVPR25/



#### **Cart-x**

# Evaluation of Multimodal Hallucinations

#### Xinya Du Assistant Professor

*University of Texas (Dallas)* https://xinyadu.github.io/

#### Liqiang Jing PhD Student

University of Texas (Dallas)



### \* Table of Content

- + What's Hallucination in MLLMs (Vision and Video LMs)
- + Causes and Interpretations of Hallucinations in MLLMs
- --- Benchmarks
- --- Metrics
  - × Reference-based
  - × Reference-free
- +- Mitigation Methods
- -- Future Work

# 1. What is Hallucination in MLLMs?

### **:** What is Hallucination in MLLMs

In the context of LVLM, the problem of hallucination can manifest as <u>textual answers</u> containing descriptions of the input <u>visual information</u> that are <u>incorrect</u>.



Hallucination in Image-related Tasks

[1] FAITHSCORE: Fine-grained Evaluations of Hallucinations in Large Vision-Language Models

### **\*** What is Hallucination in MLLMs

In the context of LVLM, the problem of hallucination can manifest as <u>textual answers</u> containing descriptions of the input <u>visual information</u> that are <u>incorrect</u>.



"Is the person drinking coffee in this video?"

"Yes, the person is drinking coffee in this video."

"Yes, the person is drinking coffee in this video."



Video-LLaVA



"Yes, these two cats are playing together."



Video-LLaVA

ShareGPT4Video

Hallucination in Video-related Tasks

[2] EventHallusion: Diagnosing Event Hallucinations in Video LLMs

### 2. Causes and Interpretations of Hallucinations

### **Causes and Interpretations of Hallucinations**



### **Hallucinations from Data Quantity**

- Deep learning models are data-hungry, especially large models like MLLMs. The amount of data plays an important role in building robust and reliable MLLMs.
- The multimodal datasets are still far less abundant than the text-only data used for training LLMs in terms of quantity.



Method		LLM	Res.	GQA	MME	MM-Vet		
InstructBLIP		14B	224	49.5	1212.8	25.6		
On	Only using a subset of InstructBLIP training data							
0	LLaVA	7B	224	-	809.6	25.5		
1	+VQA-v2	7B	224	47.0	1197.0	27.7		
2	+Format prompt	7B	224	46.8	1323.8	26.3		
3	+MLP VL connector	7B	224	47.3	1355.2	27.8		
4	+OKVQA/OCR	7B	224	50.0	1377.6	29.6		
Ad	Additional scaling							
5	+Region-level VQA	7B	224	50.3	1426.5	30.8		
6	+Scale up resolution	7B	336	51.4	1450	30.3		
7	+GQA	7B	336	62.0*	1469.2	30.7		
8	+ShareGPT	7B	336	62.0*	1510.7	31.1		
9	+Scale up LLM	13B	336	63.3*	1531.3	36.1		

[3] GRIT: Teaching MLLMs to Think with Images

### **Hallucinations from Data Quality**

- 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 synthesize 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.

#### **Hallucinations from Statistics Bias**

- Neural networks, especially large language models, possess an intrinsic tendency to memorize training data.
- The nous (e.g., objects) distribution in the training dataset has strong effects on the behavior of the model.





(a) Hallucination times of top ten frequently appearing objects, whose frequencies decrease from right to left.



(b) Hallucination times of top ten objects co-occurring with "dining table", whose frequencies decrease from right to left.

[5] Analyzing and Mitigating Object Hallucination in Large Vision-Language Models

### **Hallucinations from Language Encoder**

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

**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.

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

### **Hallucinations from Vision Encoder**

A weak vision encoder can lead to misinterpretation of visual content.



**Question**: Are there any grey points actually in the intersection of the white lines?

**Prediction**: Yes, there are grey points in the intersection of the white lines.



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.

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

### **Hallucinations from Connector**

- A weak alignment interface can easily cause hallucinations.
- The distribution gap causes trouble in cross-modal interaction, leading to hallucination



#### Original MLLM

[8] Hallucination Augmented Contrastive Learning for Multimodal Large Language Model

### Hallucinations from Inference

During generation, as the sequence length grows, the self-attention will focus more on the previously generated text tokens, i.e., the attention on the visual content is diluted.



Prompt: # Describe this image.

#### Response:

# The image depicts a busy city street with a group of people riding bicycles. There are at least 12 bicycles visible in the scene, with some of them positioned closer to the foreground and others further back.



[9] Evaluation and Analysis of Hallucination in Large Vision-Language Models

### **3. Benchmarks**



#### **Senchmarks**



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

- Dis\* means: converting the evaluation of hallucination into a binary classification task by prompting MLLMs with simple Yes-or-No short questions about the probing objects . It's not open-ended questions.
- Gen\* means generative tasks, which is open-ended questions.

[10] Hallucination of Multimodal Large Language Models: A Survey

#### \* Tasks



#### **Generative Task**

Provide a detailed description of the given image.

The image features a **table** with a variety of food items displayed in bowls. There are two bowls of food, one containing a mix of vegetables, such as **broccoli** and **carrots**, and the other containing meat. **The bowl with vegetables** is placed closer to the front, while **the meat bowl** is situated behind it. In addition to the main dishes, there is an **apple** placed on the table, adding a touch of fruit to the meal. A **bottle** can also be seen on the table, possibly containing a **beverage** or **condiment**. The table is neatly arranged, showcasing the different food items in an appetizing manner.

Land Land

#### Discriminative Task



### **4. Hallucination Metrics**



#### **K** Metrics



### **Reference-based Metric**

 Reference-based metrics evaluate the quality of generated outputs by comparing them against ground-truth references using similarity measures such as BLEU, ROUGE, or Accuracy.



$$CHAIR_{i} = \frac{|\{hallucinated objects\}|}{|\{all objects mentioned\}|}$$
$$CHAIR_{s} = \frac{|\{sentences with hallucinated object\}|}{|\{all sentences\}|}$$

CHAIR

Accuracy for POPE

### **Reference-based Metric**

 LLM-based metrics evaluate the quality of generated outputs by comparing them against ground-truth references using Large Language Models such as GPT and Gemini.

The prompt for the GPT-4 judge is designed as:

Imagine you are an intelligent teacher. Thoroughly read the question, reference answer, and the prediction answer to ensure a clear understanding of the information provided. Assess the correctness of the predictions. If the prediction answer does not conflict with the reference answer, please generate "correct". If the prediction answer conflicts with the reference answer, please generate "incorrect". If the prediction answer is unclear about the answer, please generate "unclear".

Text-Only GPT4-Assisted Evaluation in HallusionBench

### Reference-free Metric: FaithScore

 Reference-free metrics assess output quality without relying on reference texts, often using model-based scoring, rule-based heuristics, or learned quality predictors.

#### Answer

The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing...



Verifier (VEM



**Descriptive Content** The unusual aspect of this image is a man ironing clothes on the back of a minivan or van. This is not a typical place to perform this activity, as one would usually iron clothes in a more stationary and safe location, such as a home, using a regular ironing.....



#### Correlation between each evaluation metric and human judgment on LVLM hallucinations

Metric	<b>Pearson's</b> $r \%$	Spearman's $\rho$ %	Kendall's $ au$ %
BLEU-1	-15.1	-10.3	-7.5
BLEU-2	-12.7	-9.0	-6.6
BLEU-3	-7.2	-10.6	-7.6
BLEU-4	-1.9	-8.2	-5.8
ROUGE-1	-6.6	-3.0	-2.7
ROUGE-2	-5.7	-4.4	-3.4
ROUGE-L	-8.7	-6.2	-4.7
METEOR	-12.2	-8.5	-6.3
CHAIR	16.8	19.2	14.8
CLIP-Score	19.8	16.6	11.7
SPICE	20.2	21.3	25.4
Ours	48.17	38.44	47.61

[12] FAITHSCORE: Fine-grained Evaluations of Hallucinations in Large Vision-Language Models

### 5. Hallucination Mitigation

### **K** Mitigation Methods



### Control Con

- Assume that a distorted visual input would lead to text responses with more biases and priors.
- By contrasting output distributions derived from original and distorted visual inputs, VCD aims to effectively reduce the over-reliance on statistical bias and language priors



[13] Mitigating Object Hallucinations in Large Vision-Language Models through Visual Contrastive Decoding

### **High-quality Dataset: LRV-Instruction**

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



#### (Pos) Multi-Choice VQA

 $\Omega$ : Choose the correct statement about the weather conditions in the image: (a) Cloudy and rainy, (b) Clear blue sky, (c)Foggy and misty, (d) Snowy and cold A : (b) Clear blue sky



#### (Pos) Object Attribute Detection

Q: Describe the girl's hair color and whether she is wearing any accessory on her wrist. <u>A:</u>The girl's color is black and the she is wearing a bracelet on her wrist.



#### (Pos) Visual Entailment

<u>Q:</u> Verify if this statement is correct: "There is a car parking in the image." A: True



#### (Neg) Nonexistent Object Manipulation

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



#### (Neg) Knowledge Manipulation

Q: Was the youth employment rate in Ukraine gradually decreasing from 2010 to 2015? <u>A</u>: No.The youth employment rate in Morocco gradually increasing from 2010 to 2015.





(Pos) OCR <u>Q:</u> What is the number written on the lead snowboarder? A: The number is 88

Q: What objects are on toddler's feet?

(Pos) Object Detection

#### (Pos) VCR

A: White socks

<u>Q:</u> What appears to be the purpose of the green street sign? <u>A:</u> Giving directions or providing street name.

#### (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



(Neg) Knowledge Manipulation

Q: Did Merkel arrive to the Los Angeles Get Out The Vote Rally? <u>A</u>:No, Hillary Clinton arrived to the Los Angeles Get Out The Vote Rally in the image..

#### **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
<b>Finetuned MiniGPT4</b>	0.58	0.68
Original mPLUG-Owl	0.62	0.55
Finetuned mPLUG-Owl	0.69	0.78



### **Second Second S**

• Propose a new alignment algorithm called Factually Augmented RLHF that augments the reward model with additional factual information.



(c) Factually Augmented Reinforcement Learning from Human Feedback (Fact-RLHF)

[15] Aligning Large Multimodal Models with Factually Augmented RLHF

### >:< Modality Alignment: FGAIF</pre>

• Propose to align modalities in large vision-language models with Fine-Grained AI Feedback.



[16] FGAIF: Aligning Large Vision-Language Models with Fine-grained AI Feedback

### Post-processing: Woodpecker

 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.

#### Query: Describe this image.

#### MLLM Output

The scene takes place in an urban area, where a bicycle is parked next to a trash bin. The bicycle is situated on the right side of the image, with its front wheels slightly turned.

There are several other people in this area, walking and standing around at various distances from the trash bin.

#### **Key Concept Extraction**

bicycle; trash bin; person

#### **Question Formulation**

How many bicycles are there in the image?

What is next to the trash bin? Where is the trash bin located?

#### Visual Knowledge Validation

bicycle: [0.467, 0.555, 0.717, 0.746]

a recycling bin in front of the building



**Visual Claim Generation** 

#### Object-level:

 There is 1 bicycle.
 bicycle 1: [0.467, 0.555, 0.717, 0.746]

 There are 2 trash bins.
 trash bin 1: [0.145, 0.498, 0.321, 0.728];

 trash bin 2: [0.319, 0.497, 0.483, 0.729]

#### There is no person.

Attribute-level: trash bin 1: [0.145, 0.498, 0.321, 0.728] A recycling bin is next to the trash bin.

trash bin 2: [0.319, 0.497, 0.483, 0.729] A recycling bin is next to the trash bin. A bicycle is next to the trash bin.

The trash bin is located in front of the building.



The scene takes place in an urban area, where a bicycle ([0.467, 0.555, 0.117, 0.746]) is parked next to two trash bins ((0.145, 0.498, 0.321, 0.728]; [0.319, 0.497, 0.483, 0.729]). The bicycle is situated on the right side of the image, with its front wheels slightly turned. There are no people in this area.

[17] Woodpecker: Hallucination Correction for Multimodal Large Language Models

### >: Post-processing: Self-Feedback

• Utilize natural language feedback to enable the model to correct hallucinated responses by providing detailed visual information.



#### ☆ Future Work

#### --- Benchmarks

- × The lack of standardized benchmarks and evaluation metrics poses significant challenges in assessing the degree of hallucination in MLLMs
- + Cross-modal consistency issue.
- + Enhancing Interpretability and Trust.
  - × Existing methods for hallucination mitigation are primarily based on empirical observations of specific patterns. However, despite the impressive improvements achieved on specific benchmarks, understanding the underlying mechanisms and decision-making processes remains challenging.

## Thanks!

Any questions?

