

HalluciDoctor: Mitigating Hallucinatory Toxicity in Visual Instruction Data

Published at CVPR 2024, The number of citation: 53

Authors: Qifan Yu1 Juncheng Li1† Longhui Wei2 Liang Pang3 Wentao Ye1 Bosheng Qin1 Siliang Tang1 Qi Tian2 Yueting Zhuang1

Presented by - Group 5

Fazil Kagdi, Younggun Kim and Arman Sayan

Problem Statement

- **MLLMs** excel in multi-modal understanding & generation.
- **Issue:** Machine-generated instruction data causes **hallucinations**, leading to **inaccurate outputs**.

🎯 Goal of the Paper

- ✓ Detect & remove hallucinations via **cross-checking**.
- ✓ Introduce a new framework called Hallucidoctor
- ✓ Address **spurious correlations** from long-tail object co-occurrences.
- ✓ Improve **MLLM robustness** with **counterfactual data expansion**.

Types of Hallucinations in MLLMs

- 🔍 **Object Hallucination**

  - ✅ **Correct:** *"A man is sitting on a bench."*
  - ❌ **Hallucination:** *"A man is sitting on a bench next to a dog."*
  - **Dog does not exist in the image**



Types of Hallucinations in MLLMs

- 🔍 **Relation Hallucination**

  - ✅ **Correct:** *"A spoon is inside the bowl on the table."*
  - ❌ **Hallucination:** *"A cup and a spoon is on the table."*
  - **Incorrect relationship between spoon & bowl**



Types of Hallucinations in MLLMs

- 🔍 Attribute Hallucination
  - ✅ **Correct:** *"A red car is parked on the street."*
  - ❌ **Hallucination:** *"A blue car is parked on the street."*
  - **Car is actually red, not blue**



Measuring Hallucinations – CHAIR Metric

- 📊 **Standard CHAIR Metric** – Counts **nonexistent objects** in descriptions.
- 🛠️ Paper's **Extended CHAIR Metrics:**
  - **$CHAIR_{obj}$** → Measures **object hallucinations**.

$$\text{CHAIR}_{obj} = \frac{|\{\text{sentences with nonexistent object}\}|}{|\{\text{all sentences}\}|}$$

  - **$CHAIR_{rel}$** → Measures **relation hallucinations**.

$$\text{CHAIR}_{rel} = \frac{|\{\text{sentences with nonexistent relation}\}|}{|\{\text{all sentences}\}|}$$

  - **$CHAIR_{attr}$** → Measures **attribute hallucinations**.

$$\text{CHAIR}_{attri} = \frac{|\{\text{sentences with nonexistent attribute}\}|}{|\{\text{all sentences}\}|}$$

- 🚀 **Key Takeaway:** Higher CHAIR scores indicate more hallucinations.

Instance Level vs Sentence Level Hallucination

- **Hallucination Evaluation Framework:**
  - **Instance-Level:** Counts each hallucinated element (object, relation, attribute).
  - **Sentence-Level:** Evaluates if an entire sentence contains hallucinations.

🛠 POPE Metrics limitation:

✓**Limited Hallucination Scope:** Focuses mainly on object presence

Missing contextual errors

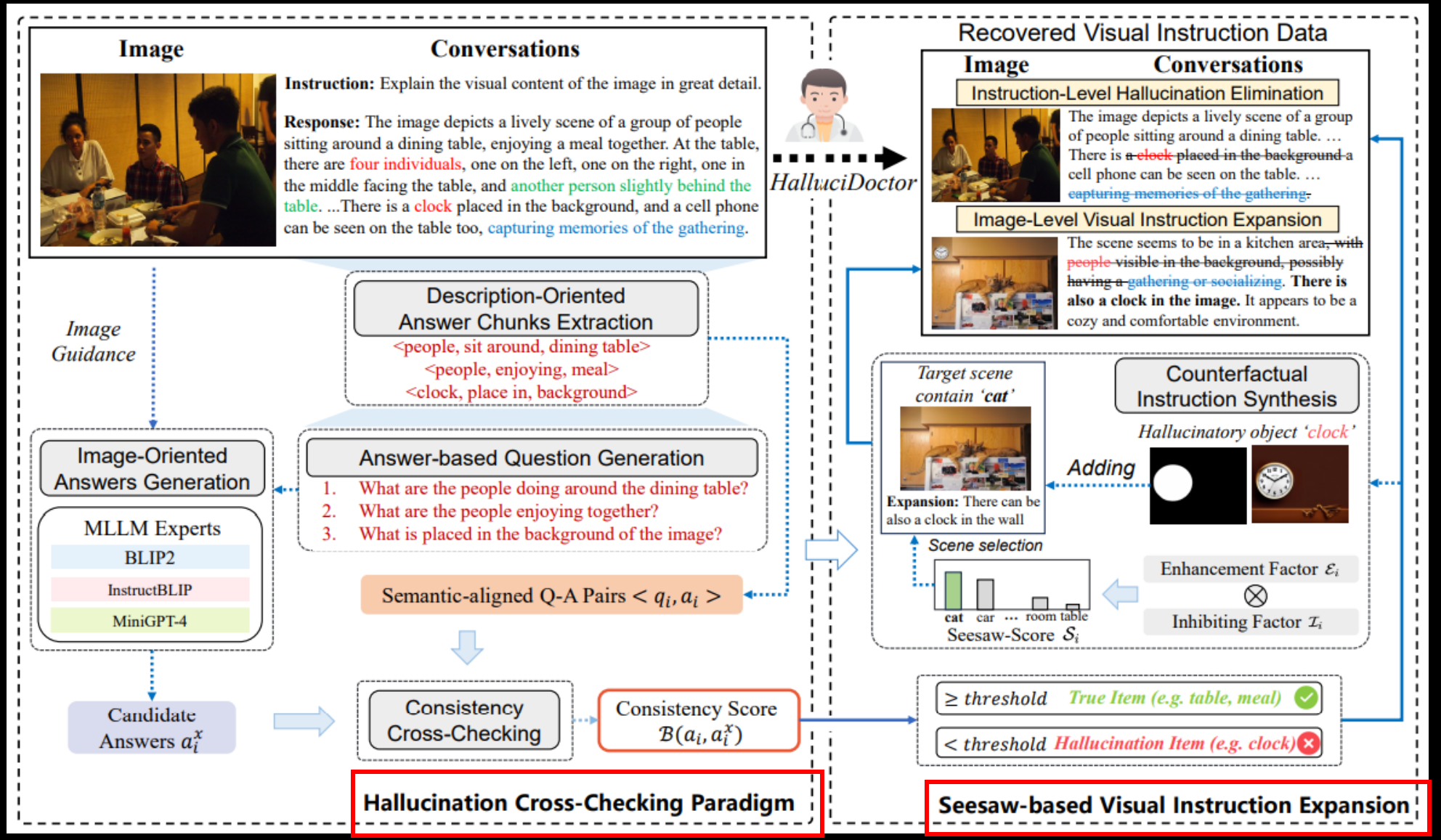
✓**Object Detection Dependency:** Relies on object detectors

Leading potential bias

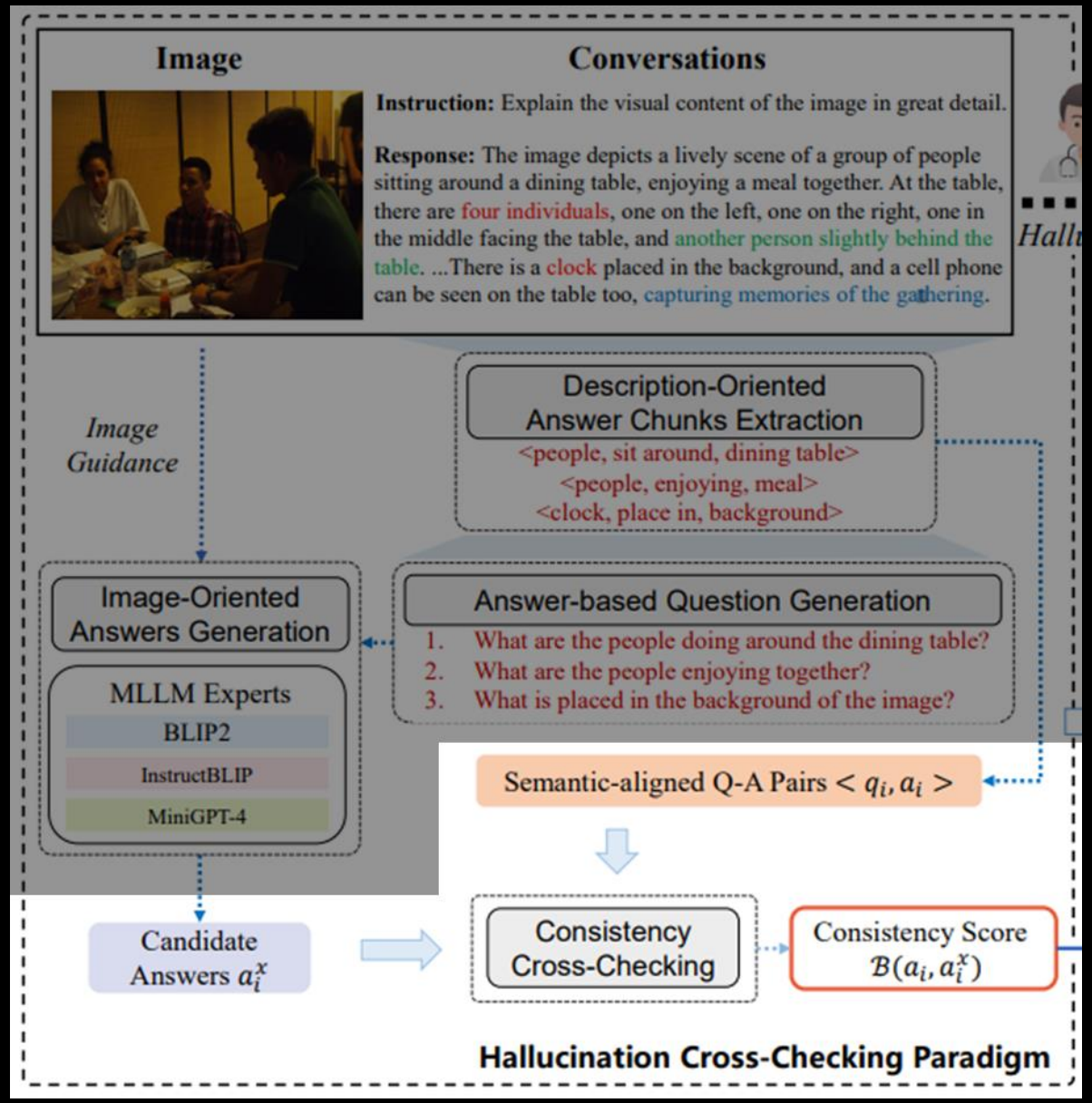
✓**Poor Attribute Evaluation:** Struggles with fine-grained attributes

e.g. color, action, relationships.

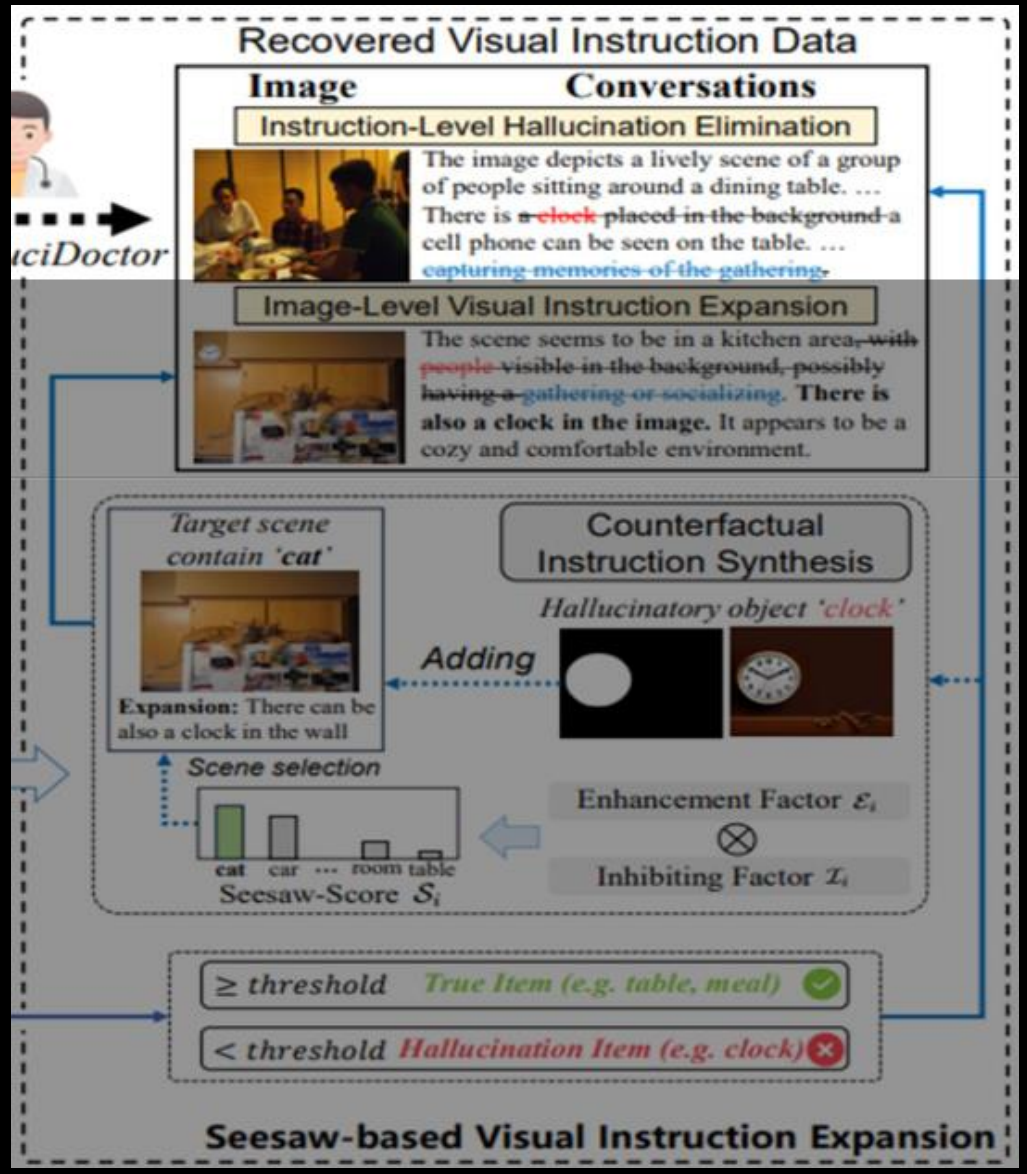
HalluciDoctor – A Two-Module Framework



Hallucination Cross-Checking Paradigm



Seesaw-based Visual Instruction Expansion



Seesaw Score

- Includes two scores
  - Enhancement Factor $\mathcal{E}_i$ (for rare objects)
  - Inhibiting Factor $\mathcal{I}_i$ (for reasonable combinations)

Task:

- Calculate $\mathcal{S}_i$
  - $\mathcal{S}_i = \mathcal{E}_i \ * \mathcal{I}_i$

Enhancement Factor

Task:

- Increase the weight of other objects o_i that rarely occur with o
- Calculate ε_i

$$\varepsilon_i = \begin{cases} \dfrac{n^*}{\max(n_i,1)}, & n_i \leq n^* \\ 1, & n_i > n^* \end{cases}$$

Enhancement Factor (contd.)

✅ **Example:**

- Hallucinated object o: car, o^*: road
- Introduce roads with objects o_i that rarely associated:

$$\varepsilon_{road-car} = \frac{1000}{1000} = 1$$

$$\varepsilon_{bicycle-car} = \frac{1000}{10} = 100$$

$$\varepsilon_{trafficLights-car} = \frac{1000}{50} = 20$$

Inhibiting Factor

Task:

- Suppress the weight of objects in rare combinations with o^*
- Calculate $\mathcal{I}_i$

$$\mathcal{I}_i = \begin{cases} \frac{m_i}{n^*}, & m_i \leq n^* \\ 1, & m_i > n^* \end{cases}$$

Inhibiting Factor (contd.)

✅ **Example:**

- Hallucinated object o: car, o^*: road
- Introduce unlikeliness for rare occurences:

$$\mathcal{I}_{car-road} = \frac{1000}{1000} = 1$$

$$\mathcal{I}_{bicycle-road} = \frac{15}{1000} = 0.015$$

$$\mathcal{I}_{trafficLights-road} = \frac{200}{1000} = 0.2$$

Seesaw Score (contd.)

✅ **Example:**

- Seesaw scores: (bicycle, traffic lights, …)

  - $\mathcal{S}_{car} = 1 \times 1 = 1$
  - $\mathcal{S}_{bicycle} = 100 \times 0.015 = 1.5$
  - $\mathcal{S}_{trafficLights} = 20 \times 0.2 = 4$

Experimental Setup and Baselines

📌 **Objective:** Evaluate **HalluciDoctor's** impact on **hallucination reduction** and **MLLM performance**.

- 🖥️ **Models Used:**
  MiniGPT-4
  mPLUG-Owl

- 📊 **Training Datasets:**
  LLaVA (Baseline)
  LLaVA+ (HalluciDoctor refined)
  LLaVA++ (Counterfactual expansion)

- 📊 **Evaluation Benchmarks:**
  CHAIR Metrics (Object, Relation, Attribute – **Instance & Sentence Level** 📖)
  MME Scores (Perception & Cognition 🧠 👀)
  GPT-4 Evaluation (Accuracy ✅ & Detailedness 📜)
  OwlEval Benchmark (Quality & Accuracy)

Chair Metrics - Hallucination Evaluation

Model Type	Methods	Instance-level			Sentence-level		
		$CHAIR_{obj} \downarrow$	$CHAIR_{rel} \downarrow$	$CHAIR_{attri} \downarrow$	$CHAIR_{obj} \downarrow$	$CHAIR_{rel} \downarrow$	$CHAIR_{attri} \downarrow$
MiniGPT4 (7B)	w/ LLaVA [25]	12.0	12.2	10.1	35.0	34.8	26.3
	w/ LLaVA+	5.9	6.1	8.5	19.6	20.5	21.9
	w/ LLaVA++	**5.7**	**5.4**	**7.1**	**16.6**	**17.1**	**20.3**
mPLUG-Owl (7B)	w/ LLaVA [25]	10.6	10.0	10.3	32.6	32.0	29.1
	w/ LLaVA+	7.6	7.1	8.0	22.2	21.1	24.2
	w/ LLaVA++	**6.4**	**5.5**	**6.7**	**19.3**	**17.6**	**16.5**

- ✓ **HalluciDoctor effectively reduced hallucination in both levels!**
- ✓ **Counterfactual samples contributed to reduce hallucination!**

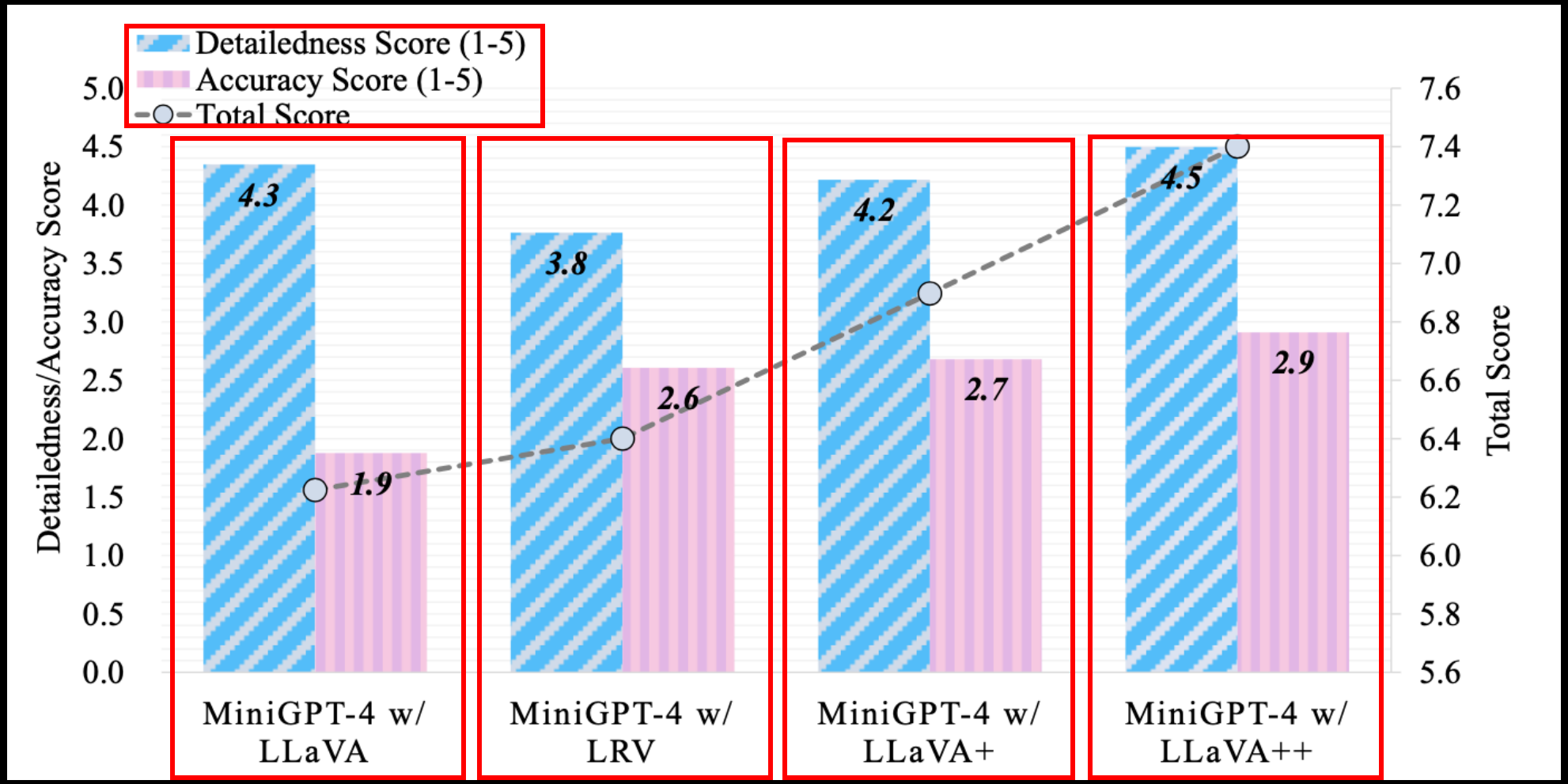
- **Intrinsic trade off between hallucination and MLLM performance**
  - ➢ **Compare not only hallucination but also MLLM performance!**

MME Benchmark to evaluate general performance

Model Type	Methods	Perception	Cognition
MiniGPT4 (7B)	w/ LLaVA [21]	859.64	289.29
	w/ LLaVA+	889.32	317.86
	w/ LLaVA++	**955.28**	**320.71**
mPLUG-Owl (7B)	w/ LLaVA	967.34	276.07
	w/ LLaVA+	1043.19	282.14
	w/ LLaVA++	**1114.34**	**302.86**

- ✓ **HalluciDoctor deleted meaningless information in datasets while maintaining semantic information**
- ✓ **Data augmentation with counterfactual samples improved perception and cognition performance**

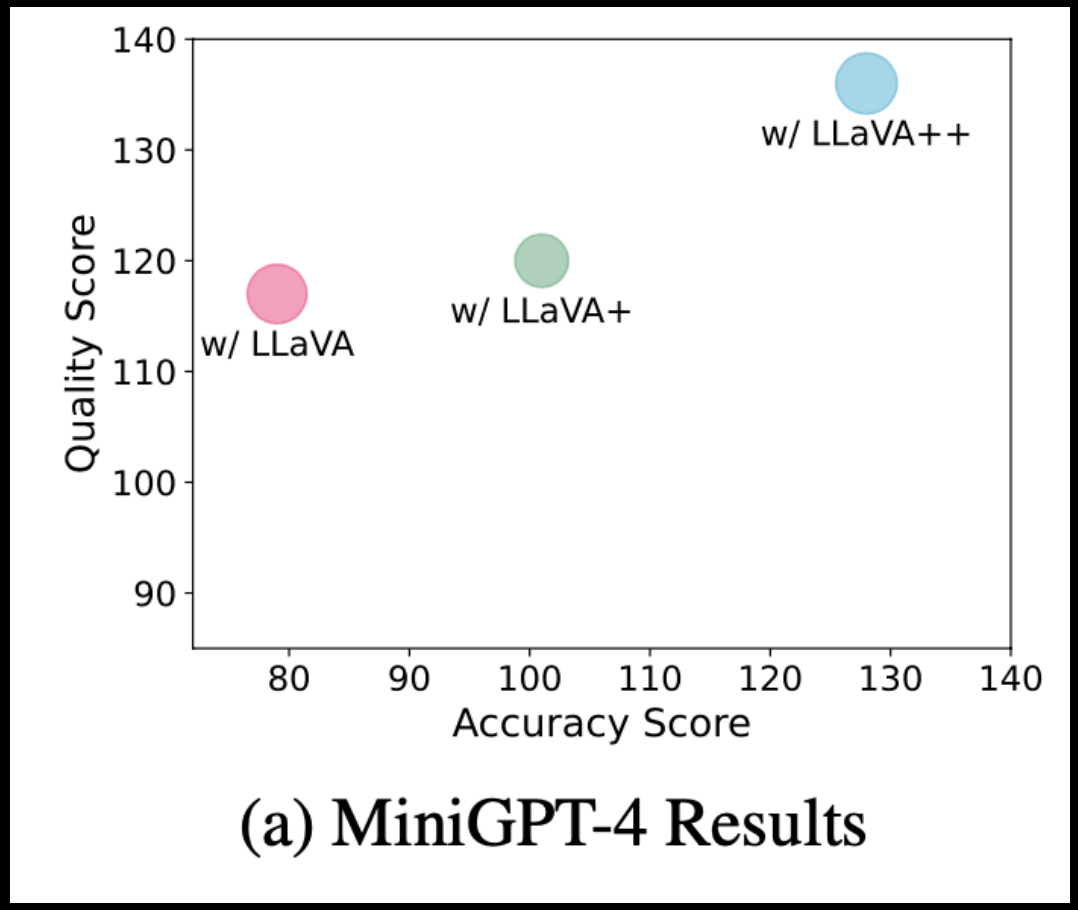
GPT-4 Evaluation – Detailedness & Accuracy



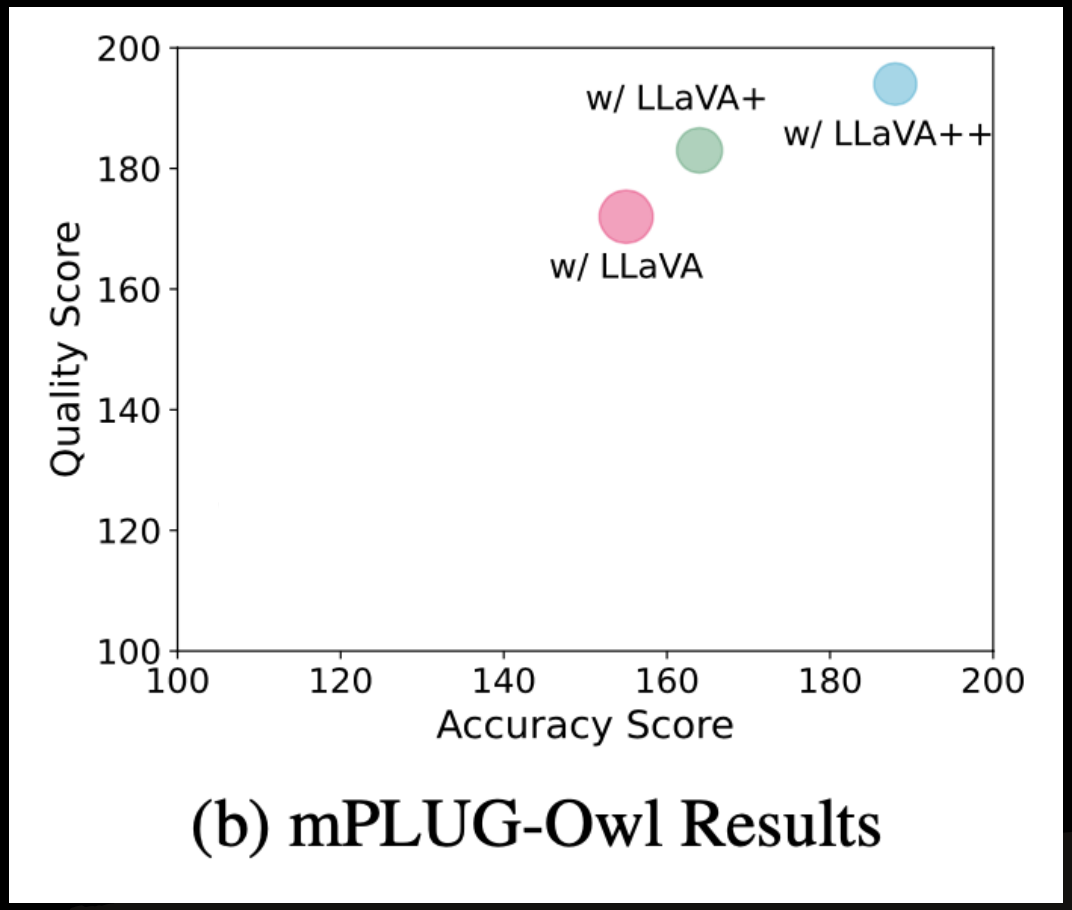
However, GPT-4 makes mistakes sometime....

OwlEval Benchmark – Human Evaluation

- ✓ **Once models generate sentences, humans manally evaluate quality and accuracy of the sentences**

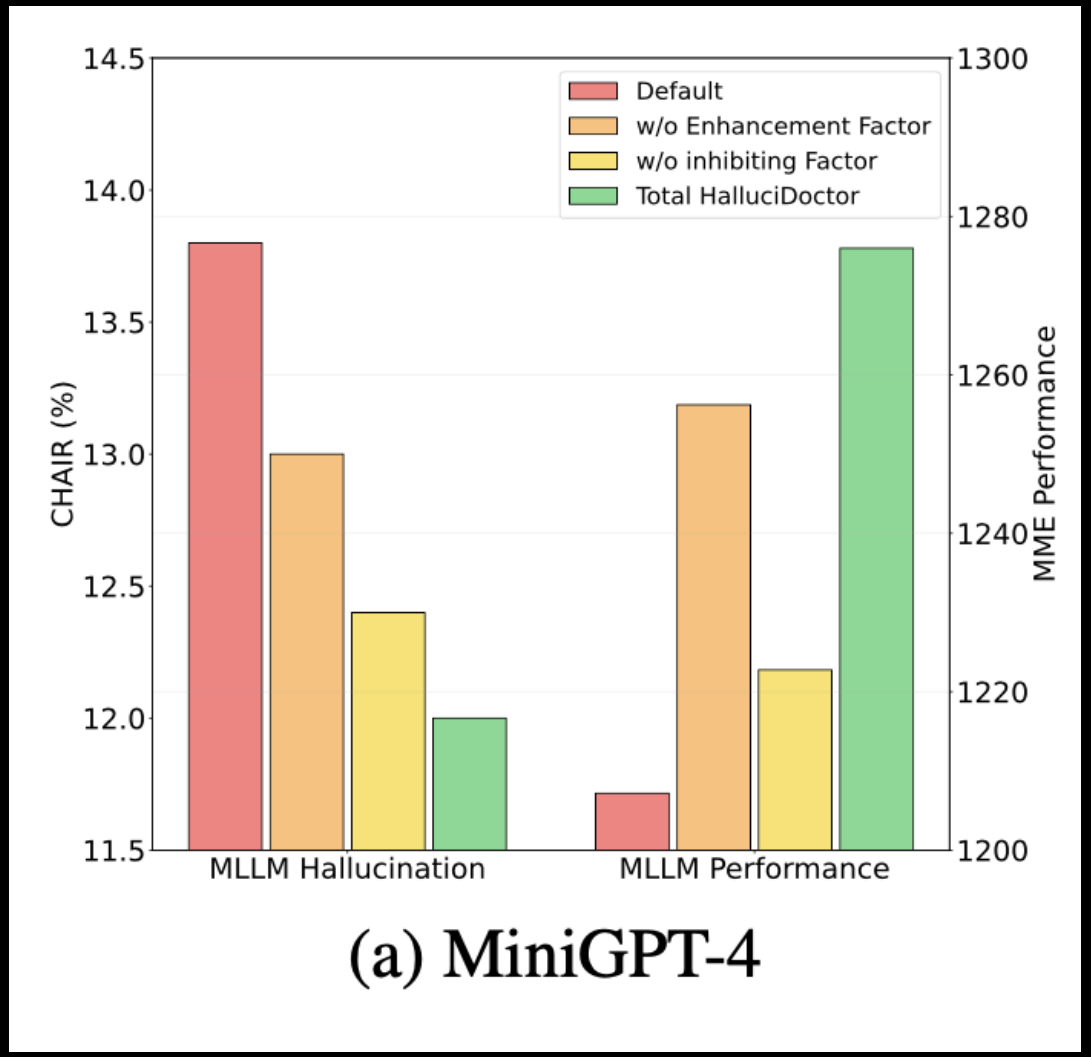


(a) MiniGPT-4 Results

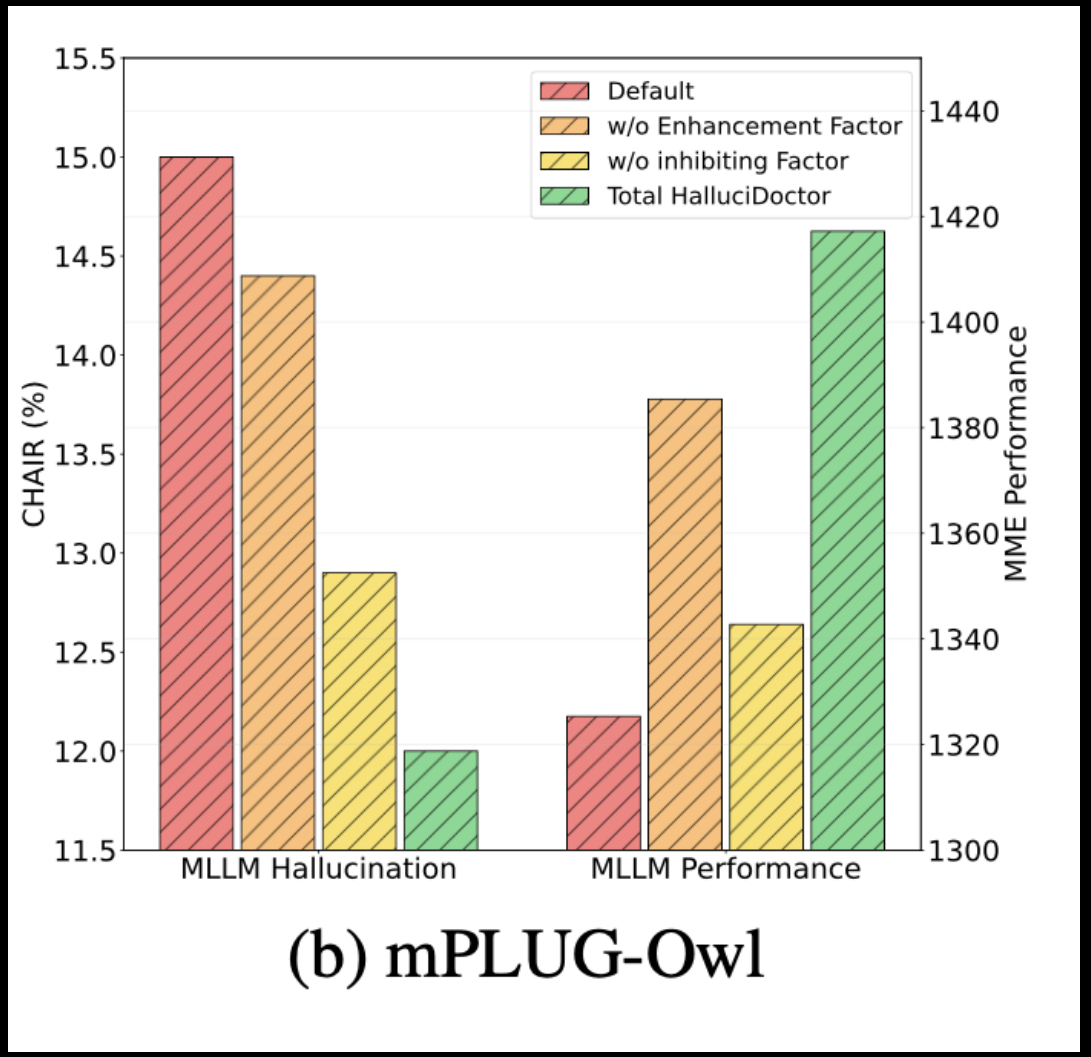


(b) mPLUG-Owl Results

Ablation Study to evaluate impact of factors



(a) MiniGPT-4



(b) mPLUG-Owl

Summary & Conclusion

✅ **HalluciDoctor's Key Benefits:**

- **Reduces hallucinations**
- **Enhances MLLM robustness against adversarial challenges**
- **Improves perception, cognition, and factual accuracy**

🚀 **Final Verdict:** HalluciDoctor **sets a new standard** for hallucination mitigation in MLLMs!

Thank You!