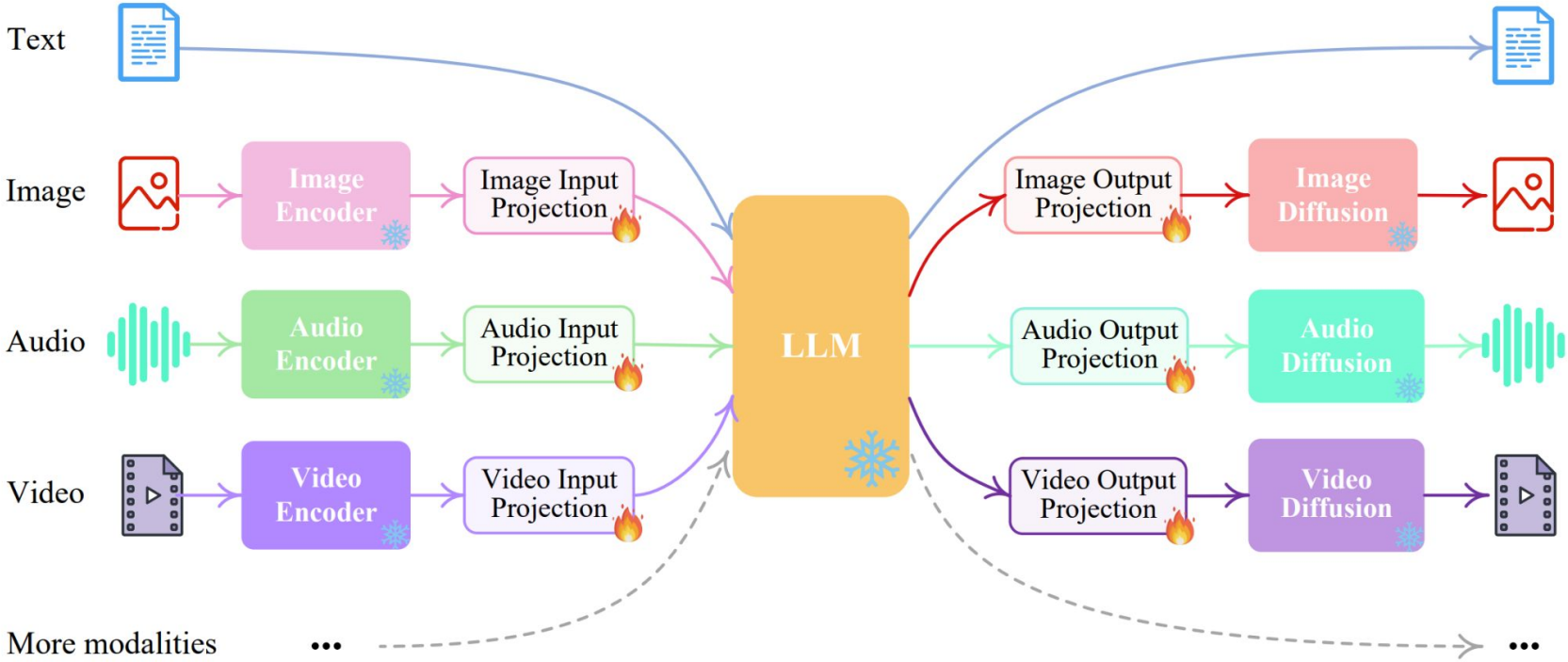
GHOST: Getting to the Bottom of Hallucinations with a Multi-round Consistency Benchmark

Vibashan VS, Nadine Chang, Jenny Schmalfuss, Vishal M. Patel, Zhiding Yu, Jose Alvarez

Multimodal Large Language Models (MLLMs)

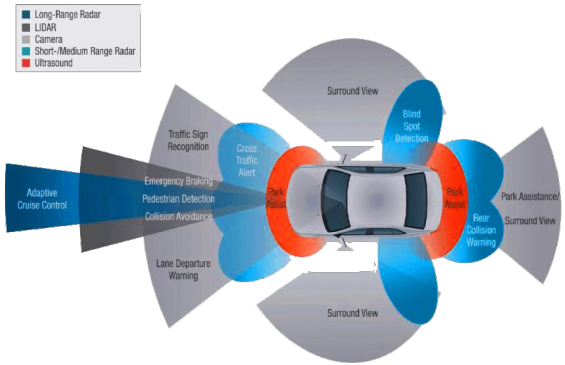


For example GPT-4o can process image, text, and audio and outputs image, text and audio.

MLLMs Application

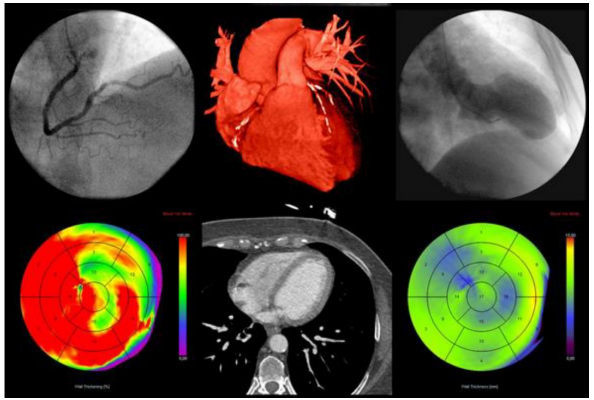
Autonomous Driving Systems

Vehicles use a combination of visual dats (cameras), spatial data (LiDAR), and auditory signals (sonar) to navigate safely.



Healthcare Diagnostics

Medical imaging tools like MRI, CT scans, and X-rays, along with patient history and verbal symptoms, are used to diagnose diseases.



Smart Home Assistants

Devices like Alexa and Google Home use voice commands (audio), physical interaction (touch), and sometimes visual cues to operate.

MLLMs are being extensively explored and applied across various domains and applications.

Hallucination in Safety Critical Applications

User: What your plan for next three seconds?

MLLM: The traffic light is green, so you can proceed once there is free space.

Hallucinations

- **Safety risks**
- **Lose Users Trust**

Definition and Common Types of Hallucinations

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

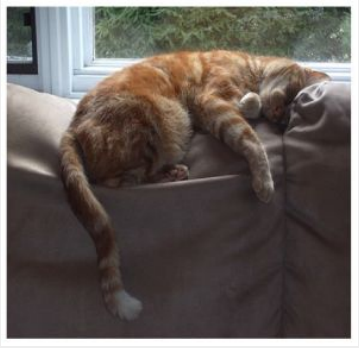
Types of Hallucination:

- **Object Hallucination:**
  - There is a shark visible in the water nearby.
  - A bird visible in the image.
- **Attribution Hallucination:**
  - The color of the surfboard is green.
  - The surfboard is made of transparent material.
- **Relation Hallucination:**
  - The man is standing on the surfboard towards the shore.
  - The man is sitting on the dog's back.

Causes of Hallucinations in MLLMs

Noisy Training Data

Pre-training stage employs image-text pairs crawled from the web, which contain inaccurate information.



Web caption: An orange cat is standing on top of the table.

Lack of Diversity

Most instruction following data samples are composed of positive instruction, without negative instruction data.

(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.

LLM Hallucination

By default the LLM hallucination is also propagated to MLLM hallucination.

How many 'm's are in the word 'Weather'?

There is one 'm' in the word 'Weather'.

Therefore, how to estimate these hallucination in existing MLLMs like ChatGPT, Gemini, etc.?

Types of Hallucination Benchmarks

Discriminative Benchmark

Use binary questions to estimate hallucination

Random settings

Is there a **bottle** in the image?

Yes, there is a bottle in the image.

Popular settings

Is there a **knife** in the image?

Yes, there is a knife in the image.

Adversarial settings

Is there a **pear** in the image?

Yes, there is a pear in the image.

Generative Benchmark

Use LLM-as-judge to estimate hallucination

Response

External LLM

Hallucinate

Not Hallucinate

Which one is more suitable for estimating hallucinations?

Types of Hallucination Benchmarks

Discriminative Benchmark

Use binary questions to estimate hallucination

Random settings

Is there a **bottle** in the image?

Yes, there is a bottle in the image.

Popular settings

Is there a **knife** in the image?

Yes, there is a knife in the image.

Adversarial settings

Is there a **pear** in the image?

Yes, there is a pear in the image.

Generative Benchmark

Use LLM-as-judge to estimate hallucination

Response

External LLM

Hallucinate

Not Hallucinate

An external LLM acting as a judge can hallucinate; therefore, direct binary questions are more effective

Existing Limitation in Hallucination Benchmarks

Image-Level Evaluation:

Object Hallucination:

User: Is there a bus in the image?

MLLM: True

Attribute Hallucination:

User: Is the color of the cyclist shirt is white?

MLLM: False

Offline:

User: Is there a color of the bus white and red?

MLLM: True

Lacks object-level evaluation: Existing benchmarks verify the presence of objects (e.g., a bus) but fail to assess specific attributes (e.g., the bus's color) or relationships between objects (e.g., the bus's position).

Existing Limitation in Hallucination Benchmarks

User: What color is the traffic light?

Sampling 1 - **MLLM:** The color of the traffic light is red.

Sampling 2 - **MLLM:** The color of the traffic light is red.

Sampling 3 - **MLLM:** The color of the traffic light is **green**.

Lacks consistency check: Stochastic nature of LLMs often leads to inconsistent answers across different iterations/sampling rounds, revealing a lack of true understanding resulting in hallucination.

GHOST Benchmark

Object-centric Evaluation

Each object (e.g., "bat") is assessed using compositional triplets consisting of the object's type, attributes, and relations.

Objects:
A **bat** is visible in the image.

Attributes:
The color of the bat is **brown.**

Relations:
The bat is to **left of** the helmet.

Consistency Check

Plausible negative variations are introduced to evaluate whether the model truly understands the object or is hallucinating.

Consistency Check Rounds

Triplet 1 ⊕	Triplet 2 ⊖	Triplet 3 ⊖
bat	ball	purse
brown	green	blue
left of	right of	in the

GHOST Object-Centric Data Collection

Object-Level Data (Bat)

	Triplet 1	Triplet 2	Triplet 3
Objects: A {**object**} is visible in the image	⊕ **bat**	⊖ **ball**	⊖ **purse**
Attributes: The color of the bat is {**attribute**}	⊕ **brown**	⊖ **green**	⊖ **blue**
Relations: The bat is {**relation**} the helmet	⊕ **left of**	⊖ **right of**	⊖ **in the**

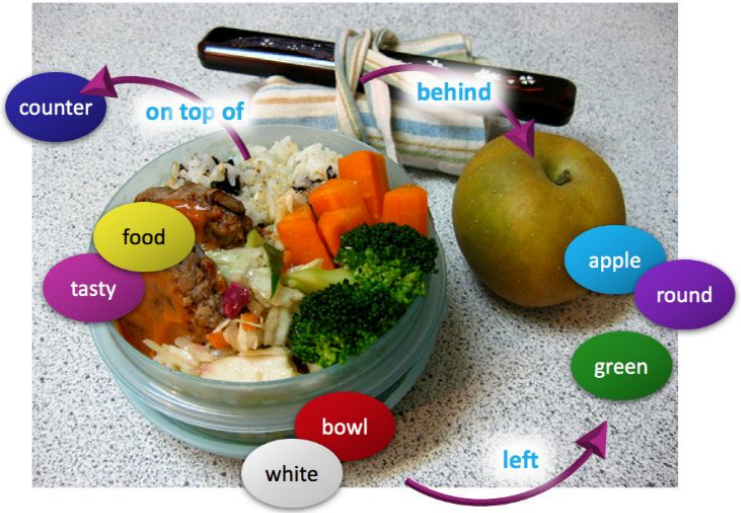
Step-by-Step Process

1. **Compositional Triplets Creation**
2. **Hard Negative Generation**
3. **Manual Filtering**
4. **Final Triplet Construction (4 triplets)**
   - Each triplet includes:
     - 1 positive (true) statement.
     - 3 hard negative (false) statements.

GHOST Object-Centric Data Collection

Step-I: Compositional Triplets Creation

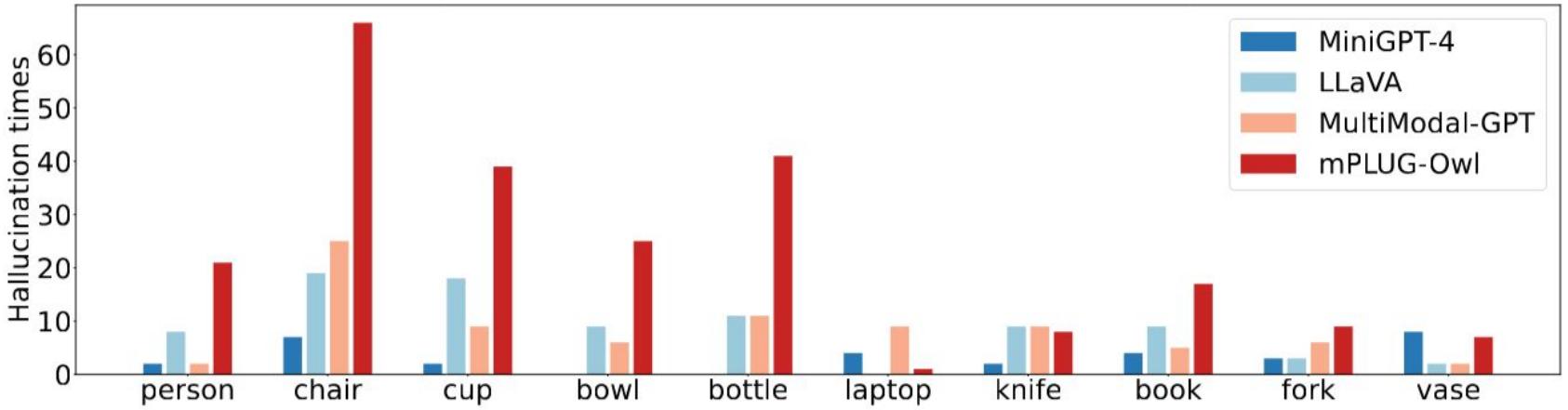
- **Dataset:** Visual Genome and GQA datasets scene graph.
- **Extraction:** Extracted objects with attributes and relations to form triplets (object type, attribute, relation).
- Example of 1 composition triplet:
  - Object: Apple
  - Attribute (Apple): Green
  - Relation (Apple, Bowl): Left



GHOST Object-Centric Data Collection

Step-II: Hard Negative Generation

- In a image, if a “dinning table” is there, then there model assume there is high chance for person, chair etc leading to hallucination.
- Motivated by this, we constructed co-occurrence matrices from datasets like Visual Genome, VQA, and LLaVA instruction Tuning datasets.
- Stratified sampling to and selected 20 plausible negatives.

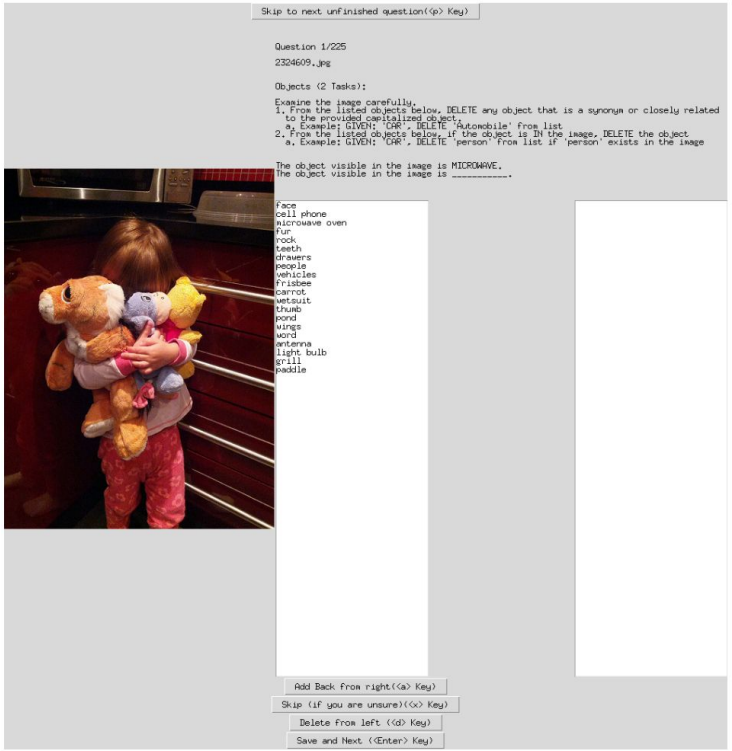


Hallucination times of top ten objects co-occurring with “dining table”, whose frequencies decrease from right to left.

GHOST Object-Centric Data Collection

Step-III: Manual Filtering

- Filtered out nonsensical or synonymous from the 20 negatives (e.g., "dog" vs. "puppy").
- Finally, selected the top three most challenging negatives for each triplet component.



GHOST Object-Centric Data Collection

Step-IV: Final Triplet Construction (4 triplets) for one object

- Each triplet includes:
  - 1 positive (true) statement.
  - 3 hard negative (false) statements.

Objects (Bat):				
A {**object**} is visible in the image.	⊕ **bat**	⊖ **ball**	⊖ **purse**	⊖ **frisbee**
Attributes (Bat):				
The color of the bat is {**attribute**}.	⊕ **brown**	⊖ **green**	⊖ **blue**	⊖ **grey**
Relations (Bat):				
The bat is {**relation**} the helmet.	⊕ **to the left of**	⊖ **to the right of**	⊖ **in the**	⊖ **above the**

GHOST Benchmark Statistics

Benchmark Statistics

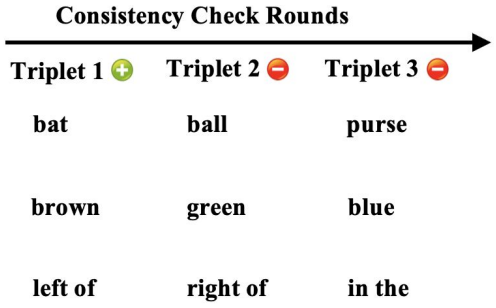
- 765 images
- 3,174 compositional triplets
- 38,088 questions
- Components:
  - 595 unique objects.
  - 385 unique attributes.
  - 81 unique relations.
- Average: 4.14 triplets per image.

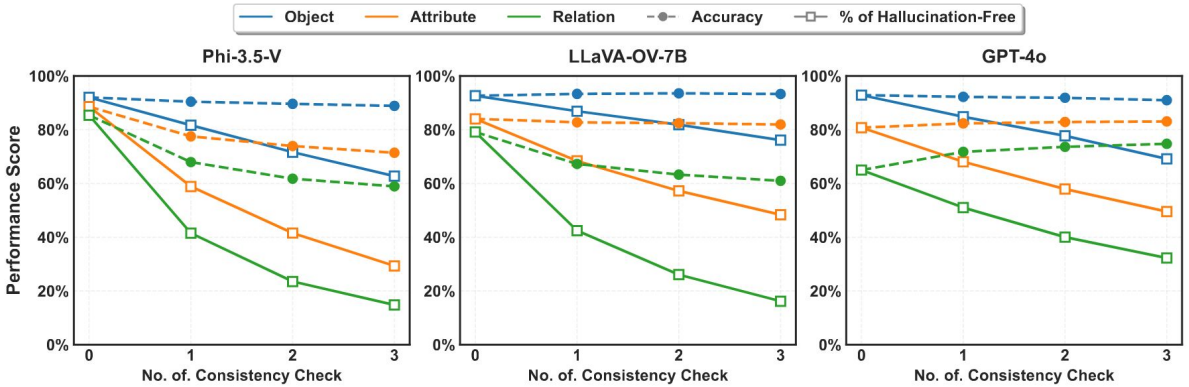
Objects:
A **stuffed bear** is present in the image.
A **vehicles** is present in the image.
A **license** plate is present in the image.
A **sign** is present in the image.

Attributes:
The color of the stuffed bear present in the image is **white**.
The color of the stuffed bear present in the image is **striped**.
The color of the stuffed bear present in the image is **pink**.
The color of the stuffed bear present in the image is **purple**.

Relations:
The spatial relation between the stuffed bear and desk is that the stuffed bear is **near** the desk.
The spatial relation between the stuffed bear and desk is that the stuffed bear is **under** the desk.
The spatial relation between the stuffed bear and desk is that the stuffed bear is **on** the desk.
The spatial relation between the stuffed bear and desk is that the stuffed bear is **of** the desk.

Consistency Check vs Accuracy

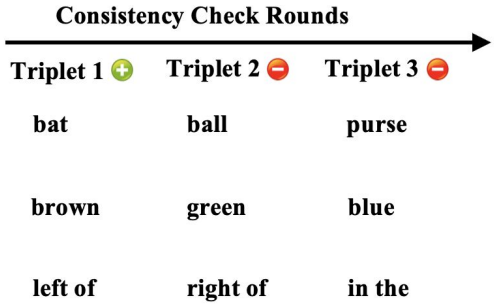


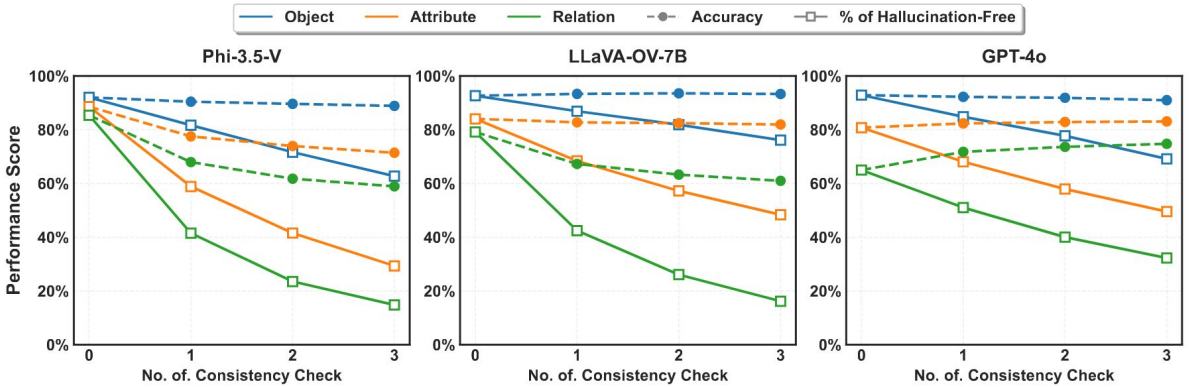


Setup: *MLLMs:* Phi-3.5-V, LLaVA-OneVision-7B, and GPT-4o | *Prompt:* Is the following statement about the image true or false: {statement}? Please answer only True or False.

Experiment: Models were tested with increasing negative triplets to assess their consistency in identifying objects, attributes, and relations. Hallucinations were flagged when models contradicted themselves within a triplet.

Consistency Check vs Accuracy





Setup: *MLLMs:* Phi-3.5-V, LLaVA-OneVision-7B, and GPT-4o | *Prompt:* Is the following statement about the image true or false: {statement}? Please answer only True or False.

Experiment: Models were tested with increasing negative triplets to assess their consistency in identifying objects, attributes, and relations. Hallucinations were flagged when models contradicted themselves within a triplet.

Conclusion: **Accuracy improves with more negatives, but hallucination-free responses drop,** showing accuracy fails to capture inconsistencies and hallucinations.

GHOST Consistency Score

GHOST Consistency Score (GCS)

Consistency Check (CC)	MLLM	brown ⊕	green ⊖	blue ⊖	Hallu-Free	Acc	GCS(CC)
Does the MLLM truly understand the image?		True	True	True	✗	33.3	14.2
<*image*> The color of the **bat** is {**attribute**}?		True	False	True	✗	66.6	42.8
1) brown 2) green 3) blue?		True	False	False	✓	100.0	100.0

$$\mathrm{GCS} = 1 - \left(\sum_{i=1}^{N_{\text{hallu}}} \frac{1}{2^{i-1}}\right) \Big/ \left(\sum_{i=1}^{N_{\text{total}}} \frac{1}{2^{i-1}}\right)$$

- GHOST Consistency Score (GCS) evaluates an MLLM's understanding by penalizing hallucinations (FP and FN) based on their frequency using a weighted geometric mean.
- GCS is a function of consistency check (CC) but accuracy is not.

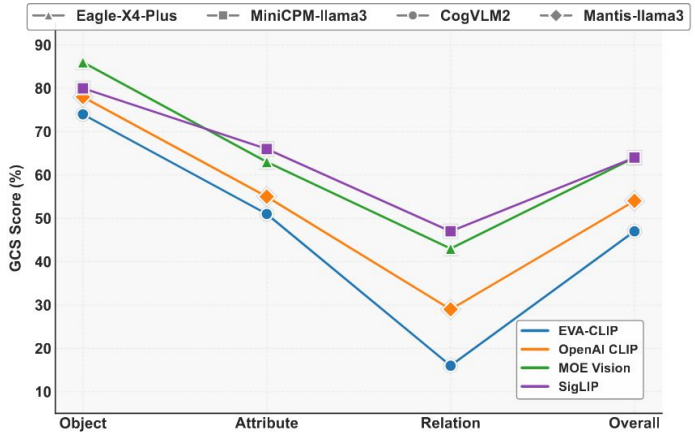
GHOST Benchmark

Model	Size	Object				Attribute				Relation				Overall			
		P	R	Acc	GCS	P	R	Acc	GCS	P	R	Acc	GCS	P	R	Acc	GCS
Tiny MLLMs (<4B)																	
LLaVA-OneVision [35]	0.5B	60.7	95.2	83.3	68.8	50.8	83.9	75.7	57.5	26.1	98.2	30.1	10.5	45.9	92.4	63.0	45.6
MiniCPM-V [54]	2B	56.2	96.1	80.3	63.9	44.0	81.8	69.8	49.8	30.1	97.7	42.7	19.8	43.4	91.8	64.2	44.5
PaliGemma [10]	3B	67.3	**96.6**	87.4	75.7	46.8	90.4	71.9	53.3	25.7	95.1	30.2	10.6	46.6	94.0	63.2	46.6
VILA-1.5 [38]	3B	71.5	93.6	89.0	78.7	55.9	77.6	79.1	62.2	29.1	95.0	41.0	18.7	52.1	88.7	69.7	53.2
Small - Medium MLLMs (4B - 13B)																	
Phi-3.5-V [4]	4B	71.6	92.0	88.9	78.3	46.3	88.6	71.4	52.2	36.3	85.4	58.9	36.9	51.4	88.7	73.1	55.8
Chameleon [47]	7B	33.2	23.4	69.1	44.7	27.3	23.0	65.5	40.8	24.5	22.3	63.4	38.8	28.4	22.9	66.0	41.4
Mantis-LLaMA3 [30]	8B	73.0	85.9	88.5	77.8	48.9	72.8	74.1	55.3	32.2	86.0	51.3	29.0	51.3	81.5	71.3	54.0
VILA-1.5 [38]	8B	84.3	83.8	92.0	84.1	57.8	73.1	79.9	63.6	31.6	62.9	56.7	33.2	57.9	73.2	76.2	60.3
Eagle-X4-Plus [46]	8B	83.9	90.2	93.2	86.2	55.6	85.6	79.3	63.1	41.1	89.8	65.2	43.1	60.2	88.5	79.2	64.1
LLaVA-OneVision [35]	8B	**82.7**	92.7	**93.3**	**86.5**	59.9	84.1	81.9	67.5	36.9	79.1	61.0	39.1	59.8	85.3	78.7	64.4
Idefics [34]	9B	46.2	94.3	71.1	50.5	41.9	67.0	68.5	47.5	28.1	84.0	42.2	19.5	38.7	81.7	60.6	39.2
LLaVA-1.5 [41]	13B	73.7	91.9	89.8	80.2	49.7	84.2	74.8	56.4	38.4	96.0	60.6	37.9	54.0	90.7	75.0	58.2
VILA-1.5 [38]	13B	81.1	89.1	92.1	84.3	**65.5**	72.1	**83.5**	68.8	38.5	83.7	62.4	40.2	61.7	81.6	79.3	64.5
Large MLLMs (> 13B)																	
MiniCPM-LLaMA3[25]	18B	73.2	92.4	89.7	79.9	58.6	80.2	80.9	65.4	38.8	47.8	68.1	46.4	56.9	73.5	79.5	63.9
CogVLM2 [22]	20B	66.3	93.4	86.5	74.0	45.5	**90.8**	70.5	51.2	28.8	**98.3**	38.8	16.1	46.9	**94.2**	65.3	47.1
InternVL-Chat [13]	26B	78.0	94.9	92.0	84.2	54.0	90.2	78.3	61.8	36.6	82.8	59.8	37.5	56.2	89.3	76.7	61.1
VILA-1.5 [38]	40B	82.3	91.4	92.9	85.9	58.1	85.0	80.9	67.7	41.0	85.0	65.7	44.4	60.5	87.1	79.8	66.0
LLaVA-OneVision [35]	72B	82.1	87.2	92.0	84.0	61.5	83.0	82.8	68.2	47.1	78.4	72.6	53.1	**63.6**	82.9	82.5	68.4
Proprietary MLLMs																	
Gemini 1.5 Pro [48]	-	71.6	94.8	89.3	79.3	59.9	83.3	81.9	67.1	41.3	77.9	66.8	46.0	57.6	85.3	79.3	64.1
GPT-4o [2]	-	76.2	92.9	91.0	82.2	62.5	80.8	83.1	**68.9**	**49.6**	65.0	**74.8**	**56.0**	62.8	79.6	**82.9**	**69.0**

GHOST Insights

Vision Encoder Effect on Hallucination

Strong vision encoders, such as SigLIP and MoE, improve GCS scores, highlighting the importance of encoder quality

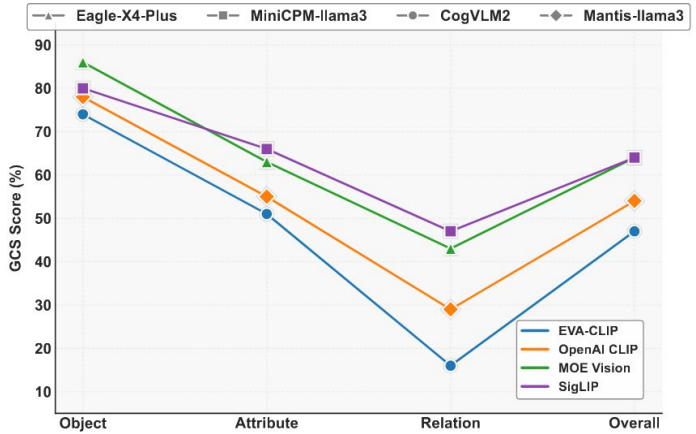


- **Strong Vision Encoder for Limited-capacity models:** Limited-capacity models with strong vision encoders are ideal for edge devices, ensuring reliability in resource-constrained and edge device applications.

GHOST Insights

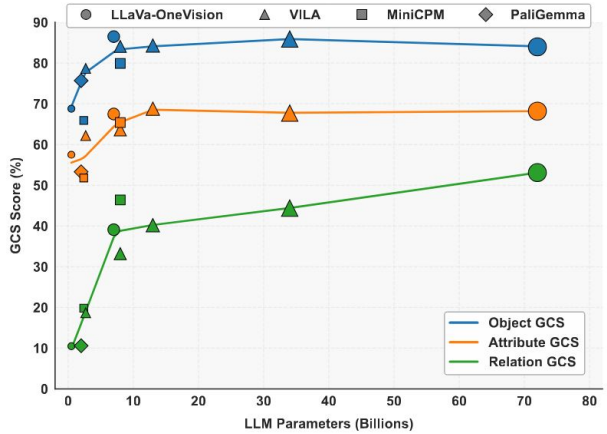
Vision Encoder Effect on Hallucination

Strong vision encoders, such as SigLIP and MoE, improve GCS scores, highlighting the importance of encoder quality



LLM Size Effect on Hallucination

Larger LLM achieve higher GCS scores, especially in relation tasks, indicating improved complex vision-language understanding with increased model capacity.



- **Strong Vision Encoder for Limited-capacity models:** Limited-capacity models with strong vision encoders are ideal for edge devices, ensuring reliability in resource-constrained and edge device applications.
- **Larger LLMs for High Accuracy:** Larger models offer better consistency and are suited for high-stakes applications like diagnostics and autonomous systems.

Summary

Object-Level Data (Bat)	Consistency Check Rounds →		
	Triplet 1 ⊕	Triplet 2 ⊖	Triplet 3 ⊖
Objects: A {**object**} is visible in the image	**bat**	**ball**	**purse**
Attributes: The color of the bat is {**attribute**}	**brown**	**green**	**blue**
Relations: The bat is {**relation**} the helmet	**left of**	**right of**	**in the**

- We present GHOST, an object-centric benchmark for evaluating hallucinations in MLLMs. GHOST offers fine-grained assessments of object types, attributes, and relations at the individual object level.
- Our novel consistency-based evaluation framework introduces the GHOST Consistency Score, a metric emphasizing false positives and negatives to better capture hallucination tendencies.
- The comprehensive dataset includes 765 images, 3,174 compositional triplets, 38,088 questions, and an average of 4.16 triplets per image, enabling detailed object-centric evaluations.
- Strong vision encoders enhance reliability in edge devices and limited-capacity models, while larger LLMs ensure high accuracy and reduce hallucinations in critical applications.

Thank You…!!!