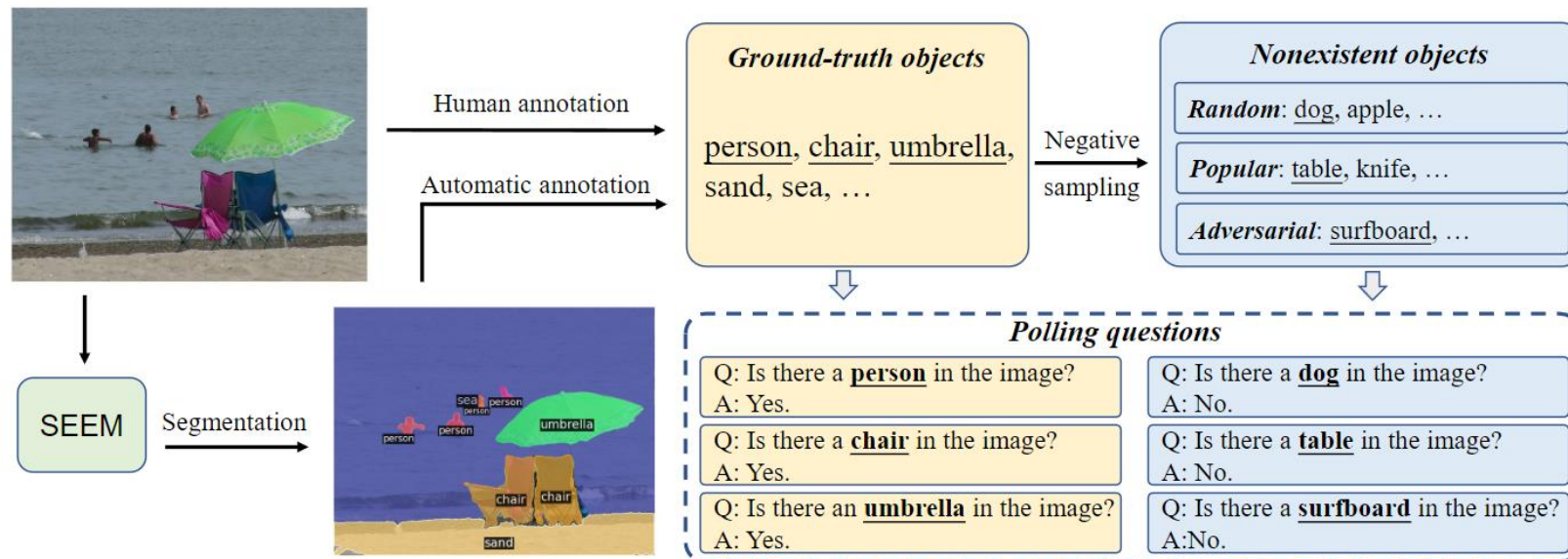


Evaluating Object Hallucination in Large Vision-Language Models

Authors: Yifan Li, Yifan Du, Zhou, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, Ji-Rong Wen

Publication: EMNLP 2023

Citations: 142



Group-6: Reeshoon Sayera, Soumik Ghosh, Ifty Rezwan, Xiao Hang Wang, Xitong Li



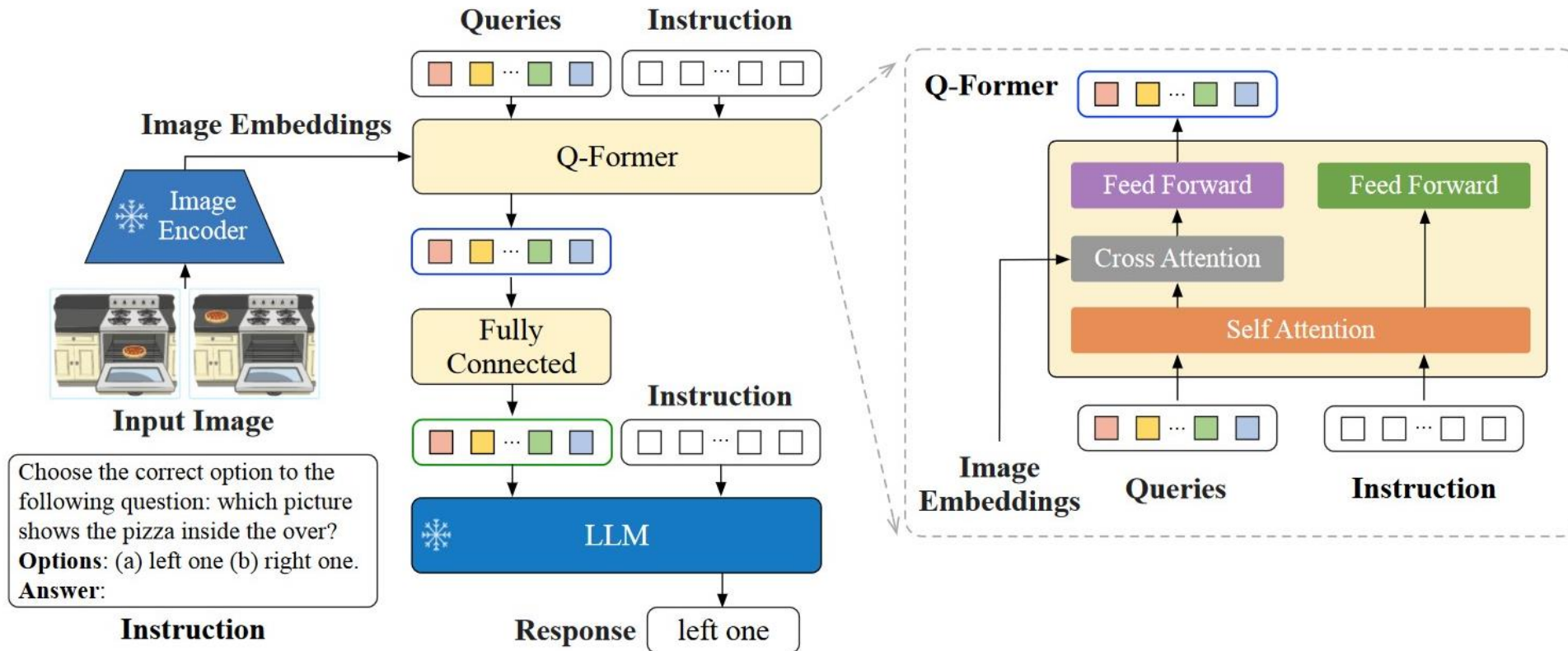
Outline

1. Introduction
2. Hallucination
3. Polling-based Object Probing Evaluation (POPE)
4. Caption Hallucination Assessment with Image Relevance (CHAIR)
5. Limitations



Introduction

- Large language models (**LLMs**) suffer from hallucination.
- Vision-language pre-trained models (**VLPMs**) also suffer from hallucination.
- Large vision-language models (**LVLMs**) = LLM + VLPM.



Hallucination

- What is **Hallucination**?

what is heavier: kilo of water or kilo of air?

A kilogram of water is heavier than a kilogram of air. Water is a dense, liquid substance that has a mass of about 1 kilogram per liter at standard temperature and pressure. Air, on the other hand, is a gas that is much less dense than water and has a lower mass. A volume of air that weighs 1 kilogram would occupy a much larger space than a liter of water, which weighs the same amount.



What is the name of the store under the green awning?

A gas station

- **NLP Definition:** *"the generated content that is nonsensical or unfaithful to the provided source content."*^[1]
- **Psychiatric definition:** *"perception-like experiences with the clarity and impact of a true perception but without the external stimulation of the relevant sensory organ."*^[2]

[1] Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., ... & Fung, P. (2023). Survey of hallucination in natural language generation. ACM Computing Surveys, 55(12), 1-38.

[2] American Psychiatric Association, 2013, p. 822

Types of Hallucination

LLMs tend to hallucinate unintended text.

The earliest mention of artificial intelligence in the New York Times was in a ~~February 19, 1950~~ ~~November 1950~~ article titled ~~“Thinking Machines:”~~ ~~“‘Revolution’ is Seen in ‘Thinking Machines.’”~~ The article, by ~~Walter Sullivan~~, reported on a meeting of the ~~American Association for the Advancement of Science~~, where a number of scientists discussed the possibility of creating machines that could think.

VLPs tend to hallucinate nonexistent objects.
(Object Hallucination)



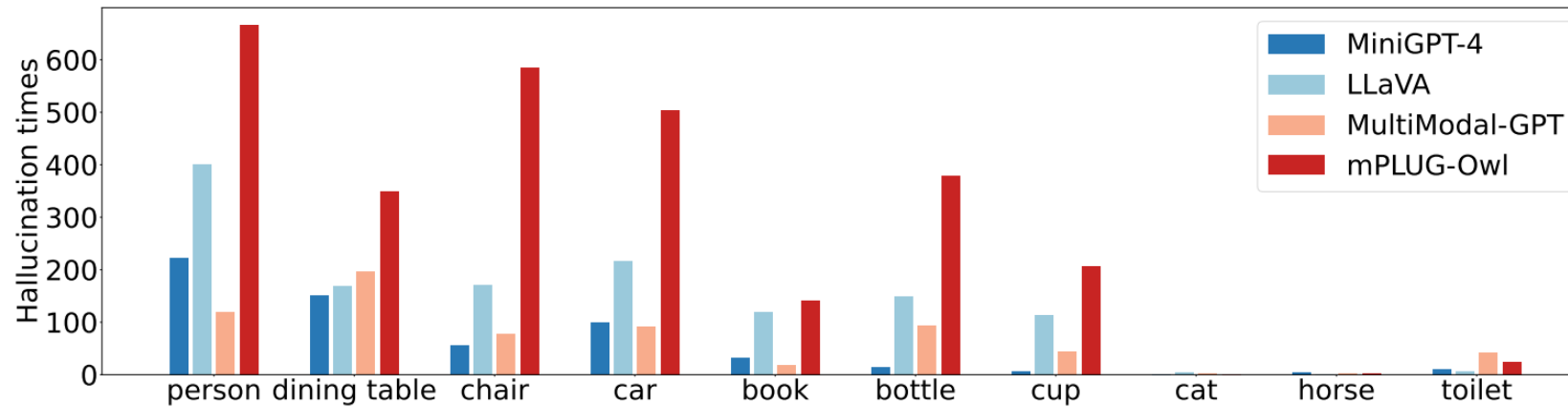
Provide a detailed description of the given image.

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.



Why does object hallucination happen?

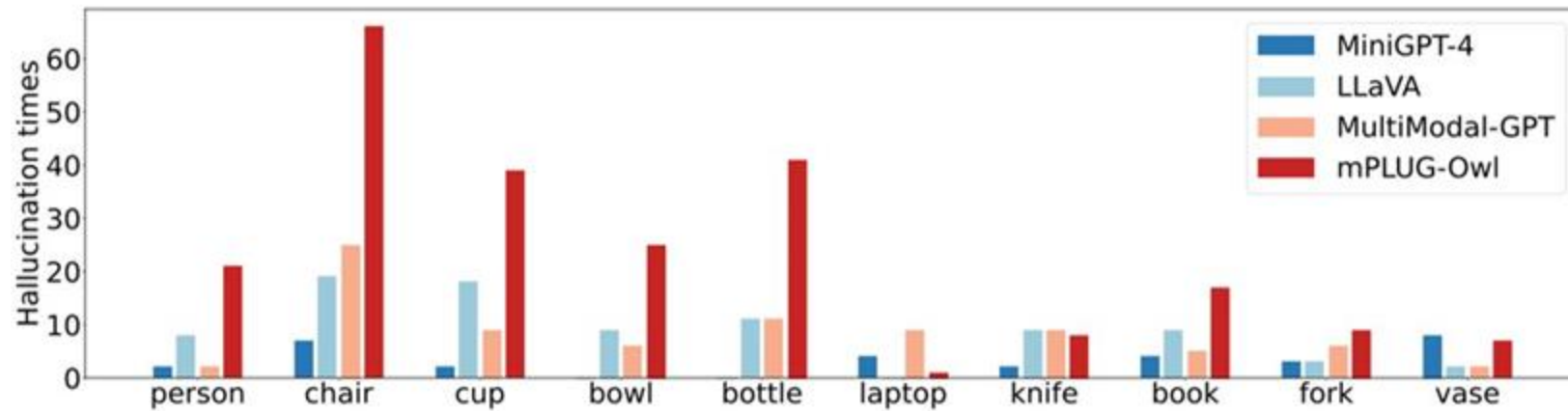
- MSCOCO has an **unbalanced object distribution** where **top frequent objects occupy a major part**.



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

Why does object hallucination happen?

- The frequently co-occurring object groups may also contribute to object hallucination.



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

Why does object hallucination happen?

$$HR_A@k = \frac{1}{n} \sum_{i=1}^n \frac{\text{Hit}@k(i)}{\text{Hallucinated}(i)}$$

$$HR_C@k(o) = \frac{1}{m} \sum_{i=1}^m \frac{\text{Hit}@k(i, o)}{\text{Hallucinated}(i)}$$

n, m : # of images.

Hallucinated(i): # of hallucinated objects in the i -th example.

Hit@ $k(i)$: # of top- k frequently appearing MSCOCO objects in **Hallucinated(i)**.

Hit@ $k(i, o)$: # of top- k frequently co-occurring objects with the probing object o in **Hallucinated(i)**.

Model	HR _A			HR _C (dining table)		
	@10	@20	@30	@10	@20	@30
mPLUG-Owl	0.5455	0.6591	0.7533	0.6608	0.7926	0.8253
LLaVA	0.4620	0.5911	0.6796	0.5628	0.7329	0.8595
MultiModal-GPT	0.4152	0.5399	0.6743	0.5742	0.7849	0.8961
MiniGPT-4	0.4610	0.5758	0.7207	0.5600	0.6980	0.9145

- ~50% of the hallucinated objects → **frequently appearing objects**.
- >50% of the hallucinated objects → **frequently co-occurring objects of "dining table"**.
- The proportion ↑ as k ↑.



Motivation: POPE



Instruction-based evaluation



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.



- Sensitivity of existing evaluation methods to instruction lengths.
- Biasedness of existing methods to short captions.
- Appearing/co-occurring frequencies of objects influence hallucination.

Polling-based Object Probing Evaluation (POPE)

POPE formulates evaluation of object hallucination as a **binary classification task**

$$\langle x, \{q(o_i), a_i\}_{i=1}^l \rangle$$



POPE

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.



Pipeline



Human annotation

Automatic annotation

Ground-truth objects

person, chair, umbrella,
sand, sea, ...

Negative
sampling

Nonexistent objects

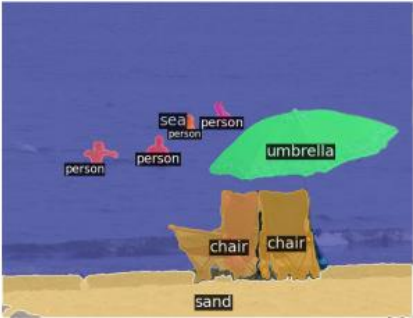
Random: dog, apple, ...

Popular: table, knife, ...

Adversarial: surfboard, ...

SEEM

Segmentation



Polling questions

Q: Is there a person in the image? A: Yes.	Q: Is there a dog in the image? A: No.
Q: Is there a chair in the image? A: Yes.	Q: Is there a table in the image? A: No.
Q: Is there an umbrella in the image? A: Yes.	Q: Is there a surfboard in the image? A: No.



Pipeline



Human annotation

Automatic annotation

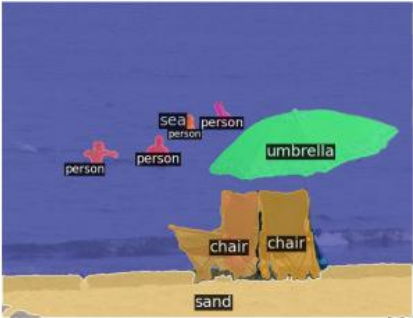
Ground-truth objects
person, chair, umbrella,
sand, sea, ...

Negative
sampling

Nonexistent objects
Random: dog, apple, ...
Popular: table, knife, ...
Adversarial: surfboard, ...

SEEM

Segmentation

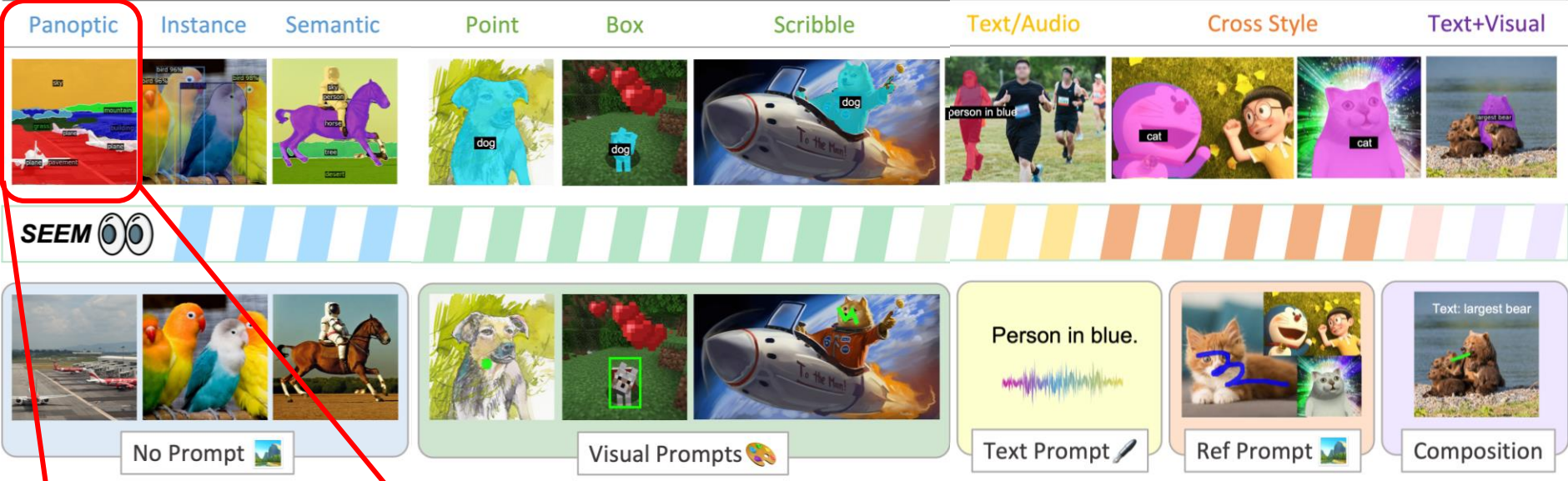


Polling questions

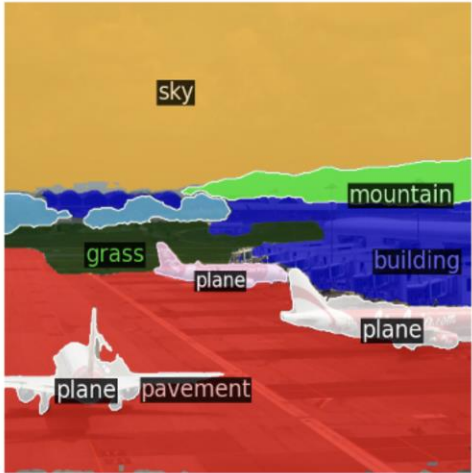
Q: Is there a person in the image? A: Yes.	Q: Is there a dog in the image? A: No.
Q: Is there a chair in the image? A: Yes.	Q: Is there a table in the image? A: No.
Q: Is there an umbrella in the image? A: Yes.	Q: Is there a surfboard in the image? A: No.



Pipeline



SEEM



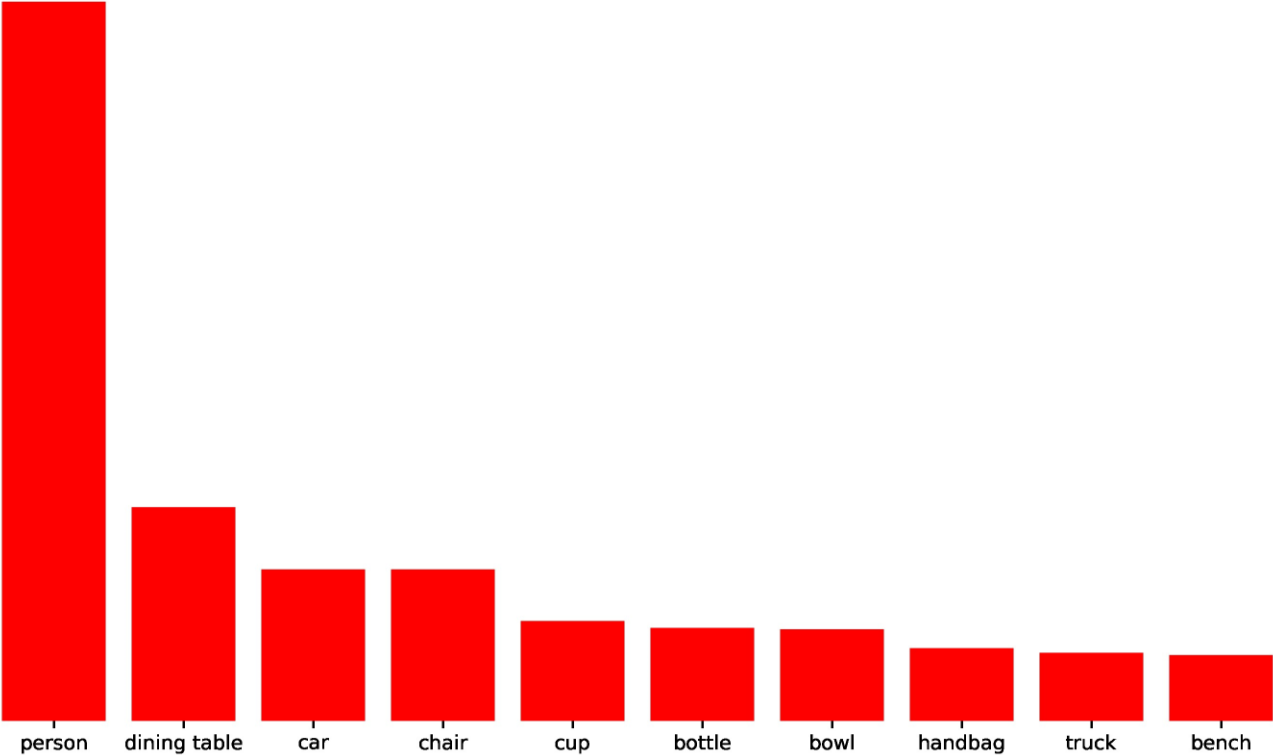
SEEM is used to get **panoptic segmentation**.

Panoptic segmentation solves both **instance segmentation** and **semantic segmentation**.



Popular Sampling

- Select top k most frequently occurring objects in the dataset but not present in the image.



Popular Sampling

- Select top k most frequently occurring objects in the dataset but not present in the image.



Ground Truth Objects

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

Answer: Yes

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

Answer: Yes

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

Answer: Yes

Non Existent Objects

Is there a **dining table** in the image?

Answer: No

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

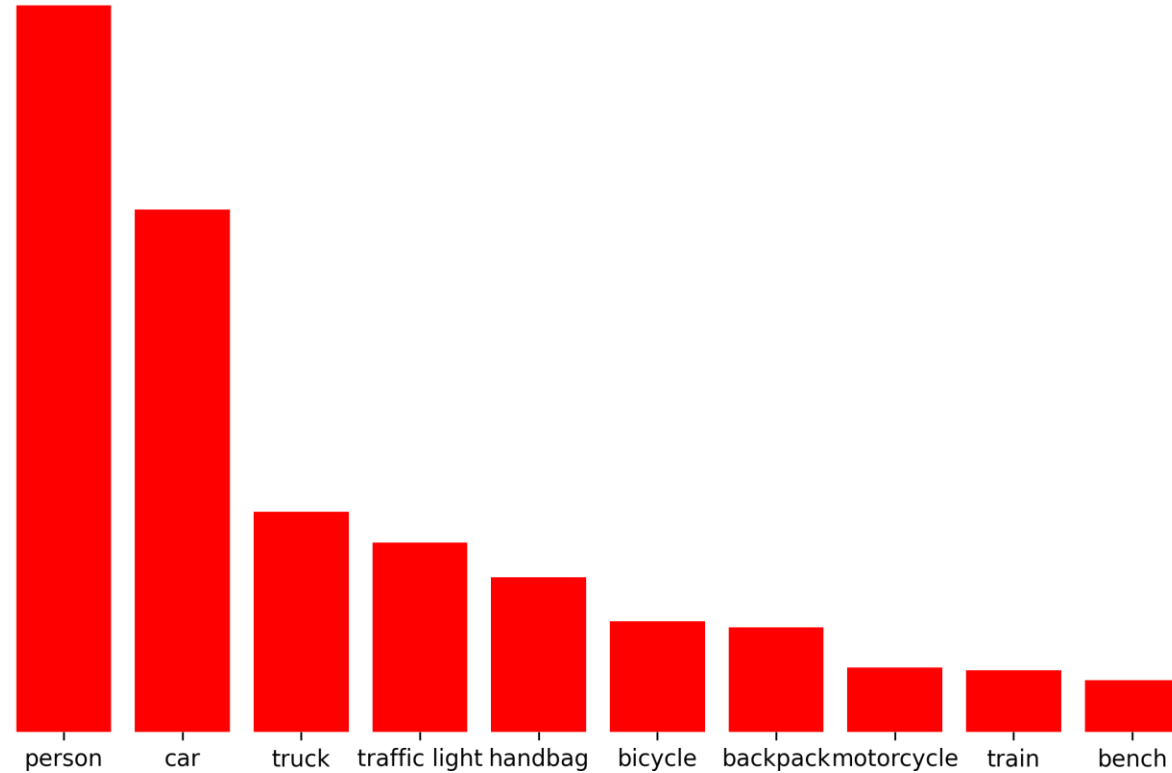
Answer: No

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

Answer: No

Adversarial Sampling

- Select top k most frequently co-occurring objects with the ground-truth.



Adversarial Sampling

- Select top k most frequently co-occurring objects with the ground-truth.



Ground Truth Objects

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

Answer: Yes

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

Answer: Yes

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

Answer: Yes

Non Existent Objects

Is there a **dining table** in the image?

Answer: No

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

Answer: No

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

Answer: No

POPE: Evaluation Settings

- **Dataset**
 - 500 images from MSCOCO val 2014
- **Question count**
 - **3 ground truth objects**
 - **3 non-existent objects**
 - **6 questions for each image**
- **Metrics**
 - Accuracy, Precision, Recall, F1 score, Yes ratio





Evaluation on MSCOCO

Difficulties: Random < Popular < Adversarial

POPE	Model	Accuracy	Precision	Recall	F1 Score	Yes (%)
<i>Random</i>	mPLUG-Owl	53.30	51.71	99.53	68.06	96.23
	LLaVA	54.43	52.32	99.80	68.65	95.37
	MultiModal-GPT	50.03	50.02	100.00	66.68	99.97
	MiniGPT-4	77.83	75.38	82.67	78.86	54.83
	InstructBLIP	88.73	85.08	93.93	89.29	55.20
<i>Popular</i>	mPLUG-Owl	50.63	50.32	99.27	66.79	98.63
	LLaVA	52.43	51.25	99.80	67.72	97.37
	MultiModal-GPT	50.00	50.00	100.00	66.67	100.00
	MiniGPT-4	68.30	64.27	82.40	72.21	64.10
	InstructBLIP	81.37	75.07	93.93	83.45	62.57
<i>Adversarial</i>	mPLUG-Owl	50.67	50.34	99.33	66.82	98.67
	LLaVA	50.77	50.39	99.87	66.98	99.10
	MultiModal-GPT	50.00	50.00	100.00	66.67	100.00
	MiniGPT-4	66.60	62.45	83.27	71.37	66.67
	InstructBLIP	74.37	67.67	93.33	78.45	68.97

Metric:

- **Accuracy** reflects the proportion of correctly answered questions.

InstructBLIP is the best due to a more diverse instruction dataset.

Evaluation on MSCOCO

Difficulties: Random < Popular < Adversarial

Metrics:

- **Precision** and **Recall** reflect the ratios of correctly answering questions whose answers are “Yes” or “No”, respectively.

POPE	Model	Accuracy	Precision	Recall	F1 Score	Yes (%)
<i>Random</i>	mPLUG-Owl	53.30	51.71	99.53	68.06	96.23
	LLaVA	54.43	52.32	99.80	68.65	95.37
	MultiModal-GPT	50.03	50.02	100.00	66.68	99.97
	MiniGPT-4	77.83	75.38	82.67	78.86	54.83
	InstructBLIP	88.73	85.08	93.93	89.29	55.20
<i>Popular</i>	mPLUG-Owl	50.63	50.32	99.27	66.79	98.63
	LLaVA	52.43	51.25	99.80	67.72	97.37
	MultiModal-GPT	50.00	50.00	100.00	66.67	100.00
	MiniGPT-4	68.30	64.27	82.40	72.21	64.10
	InstructBLIP	81.37	75.07	93.93	83.45	62.57
<i>Adversarial</i>	mPLUG-Owl	50.67	50.34	99.33	66.82	98.67
	LLaVA	50.77	50.39	99.87	66.98	99.10
	MultiModal-GPT	50.00	50.00	100.00	66.67	100.00
	MiniGPT-4	66.60	62.45	83.27	71.37	66.67
	InstructBLIP	74.37	67.67	93.33	78.45	68.97

Evaluation on MSCOCO

Difficulties: Random < Popular < Adversarial

Some LVLMs tend to answer "Yes" to every question.

POPE	Model	Accuracy	Precision	Recall	F1 Score	Yes (%)
<i>Random</i>	mPLUG-Owl	53.30	51.71	99.53	68.06	96.23
	LLaVA	54.43	52.32	99.80	68.65	95.37
	MultiModal-GPT	50.03	50.02	100.00	66.68	99.97
	MiniGPT-4	77.83	75.38	82.67	78.86	54.83
	InstructBLIP	88.73	85.08	93.93	89.29	55.20
<i>Popular</i>	mPLUG-Owl	50.63	50.32	99.27	66.79	98.63
	LLaVA	52.43	51.25	99.80	67.72	97.37
	MultiModal-GPT	50.00	50.00	100.00	66.67	100.00
	MiniGPT-4	68.30	64.27	82.40	72.21	64.10
	InstructBLIP	81.37	75.07	93.93	83.45	62.57
<i>Adversarial</i>	mPLUG-Owl	50.67	50.34	99.33	66.82	98.67
	LLaVA	50.77	50.39	99.87	66.98	99.10
	MultiModal-GPT	50.00	50.00	100.00	66.67	100.00
	MiniGPT-4	66.60	62.45	83.27	71.37	66.67
	InstructBLIP	74.37	67.67	93.33	78.45	68.97

CHAIR

- Caption **H**allucination **A**ssessment with Image **R**elevance (CHAIR)
- Calculates the proportion of objects **that appear in the caption but not in the image.**

$$\text{CHAIR}_I = \frac{|\{\text{hallucinated objects}\}|}{|\{\text{all mentioned objects}\}|}$$

Object instance level

$$\text{CHAIR}_S = \frac{|\{\text{captions with hallucinated objects}\}|}{|\{\text{all captions}\}|}$$

Sentence level



Evaluation of Object Hallucination in LVLMs with CHAIR

I	Model	CHAIR _I	CHAIR _S	Len
-	OSCAR _{Base}	7.1	13.0	-
	VinVL _{Large}	5.5	10.5	-
	OFA _{Large}	4.7	8.9	-
	BLIP _{Large}	4.7	8.8	-
I ₁	mPLUG-Owl	14.8	25.4	35.8
	LLaVA	10.5	32.7	64.3
	MultiModal-GPT	11.1	15.0	11.6
	MiniGPT-4	6.7	9.5	24.7
	InstructBLIP	2.6	3.7	8.5
I ₂	mPLUG-Owl	30.2	76.8	98.5
	LLaVA	18.8	62.7	90.7
	MultiModal-GPT	18.2	36.2	45.7
	MiniGPT-4	9.2	31.5	116.2
	InstructBLIP	2.5	3.4	7.5

I₁: "Generate a short caption of the image."

I₂: "Provide a brief description of the given image."

Len: The average length of generated captions.

- Most **instruction-tuned LVLMs suffer even more than smaller VLPMS** models.
- **Shorter prompts** give better results.

Disadvantages of CHAIR

- The evaluation results can be affected by **instruction design**.

I	Model	CHAIR _I	CHAIR _S	Len
<i>I</i> ₁	mPLUG-Owl	14.8	25.4	35.8
	LLaVA	10.5	32.7	64.3
	MultiModal-GPT	11.1	15.0	11.6
	MiniGPT-4	6.7	9.5	24.7
	InstructBLIP	2.6	3.7	8.5
<i>I</i> ₂	mPLUG-Owl	30.2	76.8	98.5
	LLaVA	18.8	62.7	90.7
	MultiModal-GPT	18.2	36.2	45.7
	MiniGPT-4	9.2	31.5	116.2
	InstructBLIP	2.5	3.4	7.5

Len: The average length of generated captions.

*I*₁: "Generate a short caption of the image."

*I*₂: "Provide a brief description of the given image."

Disadvantages of CHAIR

- CHAIR requires to examine whether the mentioned objects are hallucinated in the generated caption.

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.

Disadvantages of CHAIR

- CHAIR requires to examine whether the mentioned objects are hallucinated in the generated caption.

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.

Advantages of POPE: Stability

POPE		CHAIR	
Prompt	F1 Score	Prompt	CHAIR _I
Is there a <object> in the image?	68.65	Generate a short caption of the image.	10.50
Does the image contain a <object>?	66.83	Provide a brief description of the image.	18.80
Have you noticed a <object> in the image?	66.67	Generate a concise description for the image.	14.60
Can you see a <object> in the image?	67.58	Create a short textual summary for the image.	11.60
Avg±Std.	67.43±0.78		13.88±3.22

- **POPE** is less sensitive to instruction format.



Advantages of POPE: Scalability

Dataset	POPE	Model	Accuracy	Precision	Recall	F1 Score	F1 Score (Truth)	Yes (%)
MSCOCO	<i>Random</i>	LLaVA	50.47	50.24	99.67	66.80	68.65	99.20
		MiniGPT-4	73.77	79.25	64.40	71.06	78.86	40.63
		InstructBLIP	86.60	80.74	96.13	89.29	89.27	59.53
	<i>Popular</i>	LLaVA	50.00	50.00	99.27	66.50	67.72	99.27
		MiniGPT-4	67.80	68.80	65.13	66.92	72.21	47.33
		InstructBLIP	71.27	64.20	96.13	76.99	83.45	74.87
	<i>Adversarial</i>	LLaVA	49.77	49.88	99.20	66.38	66.98	99.43
		MiniGPT-4	61.93	61.46	64.00	62.70	71.37	52.07
		InstructBLIP	62.53	57.50	96.13	71.96	78.45	83.60

- **POPE** can be extended to unannotated datasets.

Advantages of POPE: Consistency

- Whether the **Yes/No** reflects the model's perception of objects?

Model	Number of "No" responses	Number of objects in captions
InstructBLIP	1303	0
Mini-GPT-4	1445	5

Model	Number of "Yes" responses	Number of objects in captions
InstructBLIP	664	664
Mini-GPT-4	961	1034

- Objects that get "**No**" responses from the model hardly appear in the caption.
- Models prefer to answer "**Yes**" to objects mentioned in captions.

Limitations

- Impact of hallucinations on vision tasks:
 - VQA
 - Image Captioning MSCOCO

Dataset	Model	POPE \uparrow	VQA \uparrow
A-OKVQA	InstructBLIP	87.20	59.68
	MiniGPT-4	72.47 \uparrow	38.69 \downarrow
	LLaVA	66.64 \uparrow	50.51 \downarrow
GQA	InstructBLIP	85.32	62.12
	MiniGPT-4	67.13 \uparrow	42.24 \downarrow
	LLaVA	66.56 \uparrow	47.60 \downarrow

Model	POPE	BLEU-1	BLEU-2	METEOR	ROUGE-L
InstructBLIP	89.29	59.5	45.2	22.6	42.3
LLaVA	68.65	22.0	13.9	19.8	22.4
MiniGPT-4	78.86	41.1	28.8	25.6	44.7

Limitations

- Limited computational resources may cause **partial evaluations on validation sets**, potentially skewing results due to data distribution variations.
- Automatic segmentation tools can lead to **discrepancies in object annotations** compared to human annotations, impacting evaluation results due to inconsistent label sets.
- POPE's matching-based approach for LVLM response determination might result in inaccuracies when **LVLMs fail to explicitly include predefined keywords**.
- Only a **small number of LVLMs have been evaluated** with POPE.



Potential Improvements

- A **weighted average** of metrics.
- Instead of evaluating on the part of the validation dataset, we **overfit a model** and find a way **evaluate the distribution produced** by the model.
- Leverage **task-based segmentation models** for **evaluation** so the divergence occurs less.
- Instead of an explicit 'Yes' or 'No' Question-Answer, we make the **metric flexible** by evaluating on **top-k evaluations**.
- **Create an arena and leaderboard** like **LMSYS Arena Chat** where **authors** can **upload their models** and **evaluate** them.



Thank you.

