CAP 6412 - Paper Review & Analysis

Date: 02/19/2025

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

Authors: Leng, S. et al. (2023)

Venue: CVPR, 2024

Citations: 151

Paper URL: https://arxiv.org/pdf/2311.16922



Presenters

Inad Alqurashi, Sabbir Mollah, Ahnaf Munir, Qingyang Liu, Saurabh Aggarwal

Advisor:

Dr. Mubarak Shah

Presentation Outline

- Object Hallucination?
- Causes of Object Hallucination
- Proposed Method:
  - Visual Contrastive Decoding (VCD)
  - Adaptive Plausibility Constraints
- Experimental Setup & Results
- Ablation Studies & Discussion
- Limitations & Conclusion

Object Hallucination

Hallucination in LLM

LLMs can generate text that sounds plausible but is incorrect.

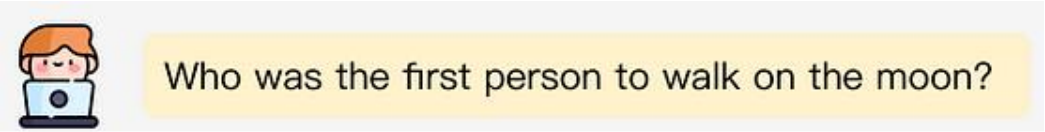


Figure from: https://vivedhaelango.substack.com/p/how-to-detect-hallucinations-in-llm?utm_campaign=post&utm_medium=web

Object Hallucination in LVLM

Object Hallucination is when LVLMs generate descriptions that include objects not present in the image.

Prompt: Describe this photo in detail.

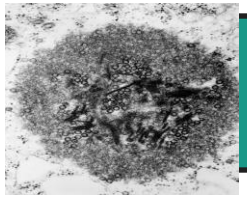
Regular Decoding: The image features a painting of a brown dog sitting in a chair, with a turkey feather in its mouth, giving it a unique, conspicuous appearance. The dog's expression is one of mischief as it looks directly into the camera. The scene is set in a room with a **dining table** and a **fork** next to the **table**. A rock, likely from the surrounding landscape, can be seen beside the **table**. A **person**'s face is visible near the bottom left corner of the painting, partially covered by the dog's furry coat.



Figure from: https://arxiv.org/pdf/2311.16922

Why Hallucination Matters?

Hallucination can mislead decision making in **healthcare**, **autonomous driving**, **robotics,** etc.



Misdiagnosing a tumor



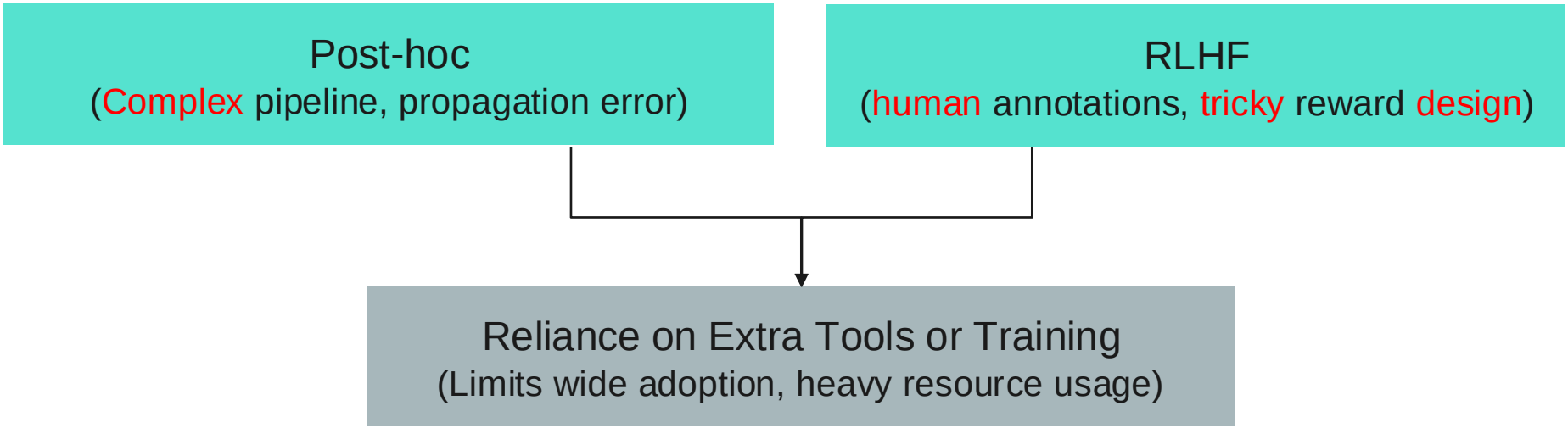
Car that suddenly stops after hallucinating a stop sign



Trying to put materials into a non-empty shelf

The Is Photo By Photo Author Is Licensed Under CCYYSA.

Existing Solutions & Limitations



What Causes Object Hallucination

Primary Causes of Object Hallucination



Statistical bias



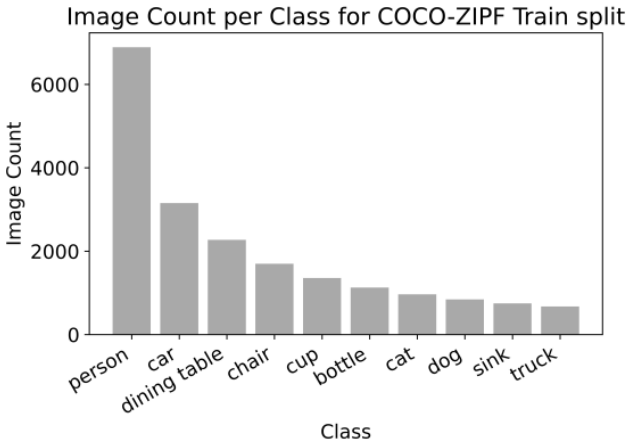
Language priors



Visual Distortion

Statistical Bias

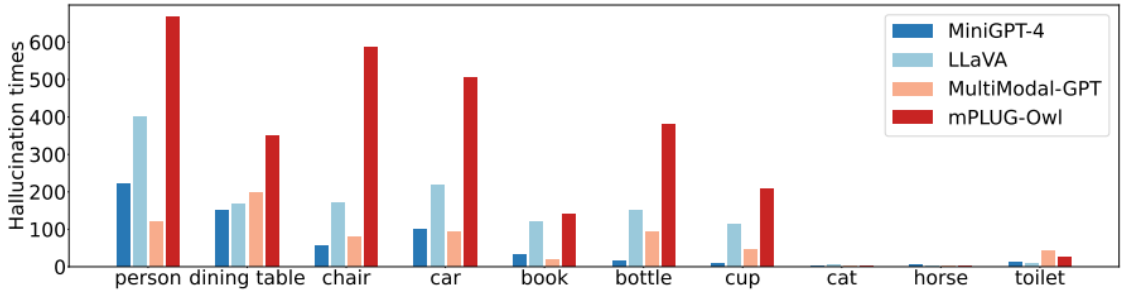
- Frequent objects overshadow actual visual evidence
- Frequently co-occurring objects can also cause bias



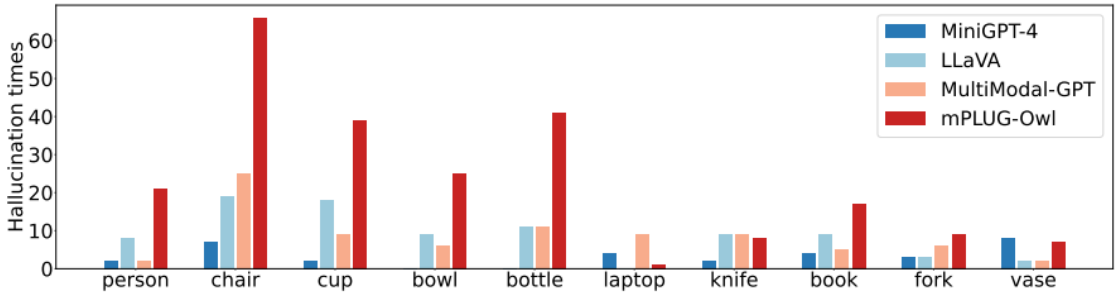
Top co-occurring objects with Dining Table:
Person, Cup, Bottle, ...



Statistical Bias



Hallucination of frequent objects

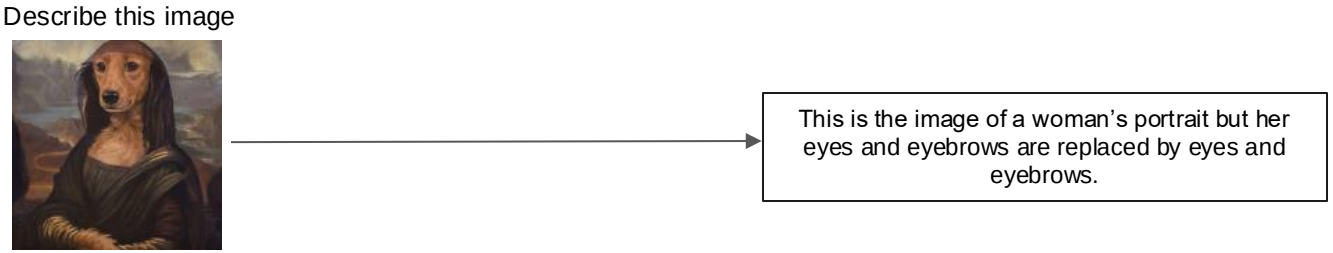


Hallucination of co-occurring objects with “DINING TABLE”.

Figures from: https://arxiv.org/pdf/2305.10355

Language Priors

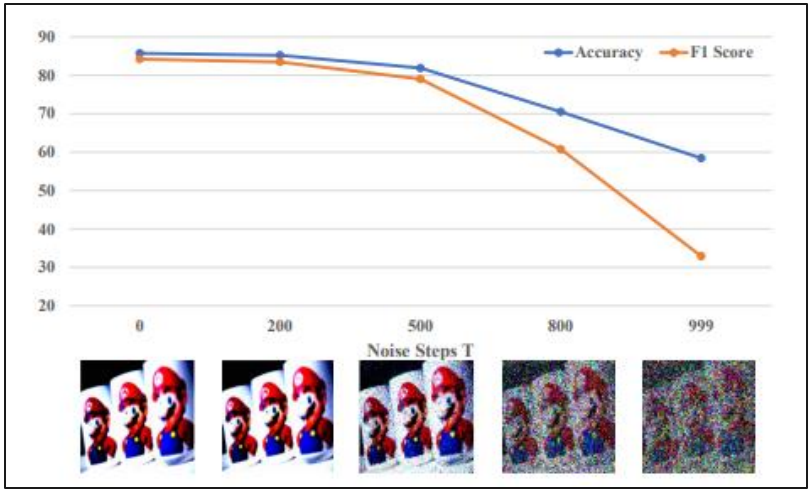
- VLMs may rely excessively on language patterns.
- Common phrases dominate, even when inconsistent with the image.



Figures from: https://arxiv.org/pdf/2311.16922

Visual Distortion

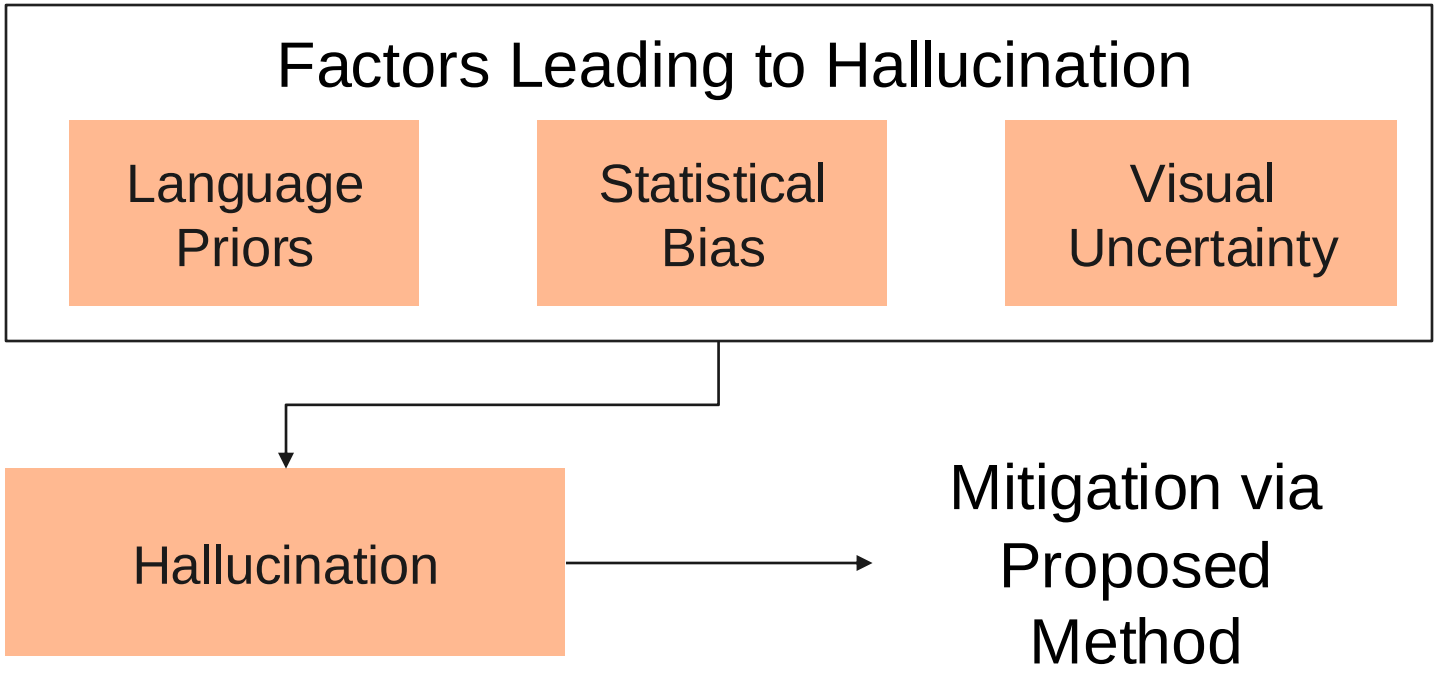
- Visual uncertainty (distortion) amplifies statistical bias and language priors.
- Validation: Incrementally adding noise gradually degrades LVLM performance.



LLaVA-1.5's performance on the POPE benchmark

Figures from: https://arxiv.org/pdf/2311.16922

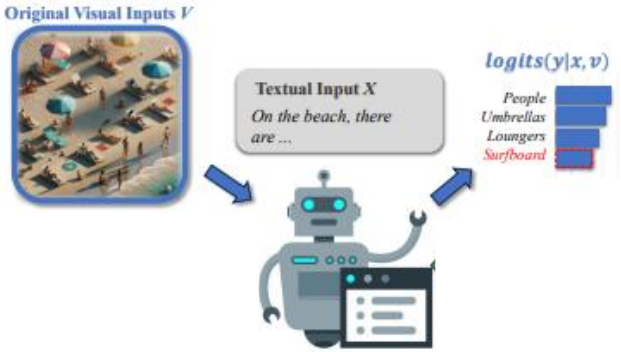
Synergy


Factors Leading to Hallucination
Language Priors
Statistical Bias
Visual Uncertainty
Hallucination
Mitigation via Proposed Method

Proposed Method

Visual Contrastive Decoding (VCD)

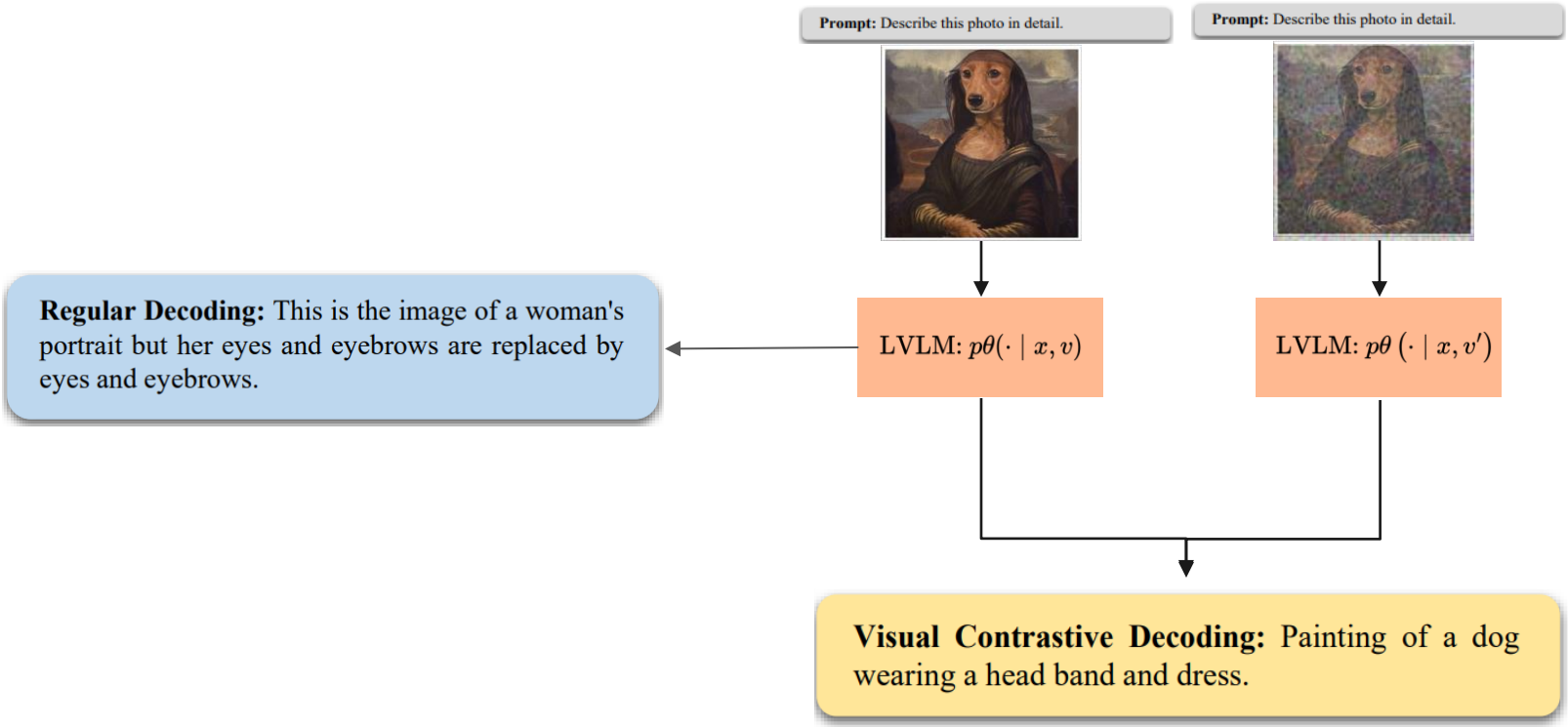
Proposed Method – Visual Contrastive Decoding (VCD)



- Produce logits from the **original** visual input
- Produce logits from the **distorted** visual input
- Subtract the distorted logits from the original logits

Figure from: https://arxiv.org/pdf/2311.16922

VCD in Action



Figures from: https://arxiv.org/pdf/2311.16922

Another intuition for VCD

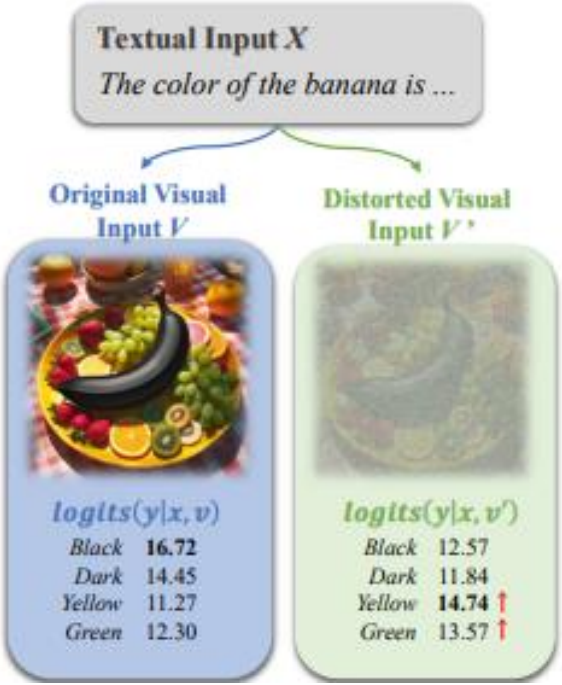


Figure from: https://arxiv.org/pdf/2311.16922

Understanding VCD Mathematically

$$p_{vcd}(y \mid v, v', x) = softmax[(1+\alpha)\, logit_\theta(y \mid v, x) - \alpha\, logit_\theta(y \mid v', x)]$$

Logits for actual image

Logits for distorted image

Next Token

Original image

Distorted image

Text

Equation from: https://arxiv.org/pdf/2311.16922

Role of α (Contrast Strength)

$$p_{vcd}(y \mid v, v', x) = softmax[(1+\alpha)\, logit_\theta(y \mid v, x) - \alpha\, logit_\theta(y \mid v', x)]$$

- α = 0 disables contrastive decoding;
- More about tuning this parameter will be explained in the ablation studies

Equation from: https://arxiv.org/pdf/2311.16922

Pros of VCD

- Penalizes hallucination-prone tokens.
- Grounds output to visual content.
- Offers a training-free method avoiding extra data or models.
- Lowers over-reliance on language priors.
- Maintains minimal computational overhead.

Proposed Method

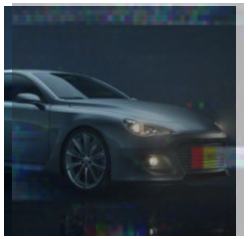
Adaptive Plausibility Constraints

A challenge in VCD!

- Penalizes entire model's output based on distorted image's results.
- Indiscriminate penalization of correct tokens might lower the performance.
- Distorted visual inputs do not always lead to incorrect responses.
- VCD can introduce improbable tokens

Intuition for Adaptive Plausibility Constraints





After VCD

TOKEN	Logit
car	0.70
cat	0.20
dog	0.03
...	...

-

TOKEN	Logit
car	0.60
cat	0.10
dog	0.02
...	...

=

TOKEN	Logit
car	0.10
cat	0.10
dog	0.01
...	...

Equation from: https://arxiv.org/pdf/2311.16922

Adaptive Plausibility Constraints

- Create a set of tokens that have a high enough confidence with original image

$$V_{head}(y_{<t}) = \{y_t \in V : p_\theta(y_t | v, x, y_{<t}) \geq \beta \max_w p_\theta(w | v, x, y_{<t})\}$$

- Probabilities of tokens not in the set are set to 0

$$p_{vcd}(y_t | v, v', x) = 0, \qquad if\, y_t \notin V_{head}(y_{<t})$$

- Ensures that VCD does not introduce improbable tokens

Equation from: https://arxiv.org/pdf/2311.16922

Combining VCD with Adaptive Plausibility Constraint

$$y \sim softmax[(1+\alpha)\, logit_\theta(y_t|v, x, y_{<t}) \; - \; \alpha\, logit_\theta(y_t|\, v', x, y_{<t})],$$

subject to $y_t \in V_{head}(y_{<t})$

Equation from: https://arxiv.org/pdf/2311.16922

Experimental Setup & Results

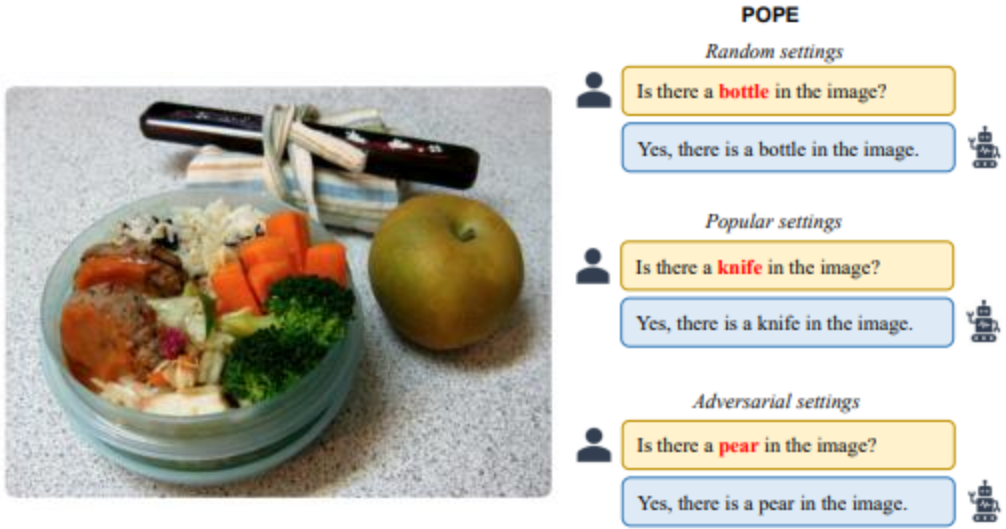
Datasets & Metrics

- Datasets:
  - POPE (Polling-based Object Probing Evaluation)
  - MME (covering perception and cognition tasks)
  - LLaVA-Bench (open-ended vision–language testing)

- Metrics:
  - Accuracy, Precision, Recall, F1 for POPE
  - Accuracy/Accuracy+ in MME’s official framework
  - GPT-4V-based Accuracy and Detailedness on LLaVA-Bench

What is POPE

- POPE benchmarks how well an LVLM discerns if a specified object is present or absent, capturing its tendency to hallucinate objects.



Equation from: https://arxiv.org/pdf/2311.16922

POPE Results

- VCD outperforms baselines in random, popular, and adversarial tasks.

Dataset	Setting	Model	Decoding	Accuracy↑	Precision	Recall	F1 Score↑
MSCOCO	*Random*	LLaVA1.5	Regular	83.29 (±0.35)	92.13 (±0.54)	72.80 (±0.57)	81.33 (±0.41)
			VCD	**87.73** (±0.40)	91.42 (±0.55)	83.28 (±0.42)	**87.16** (±0.41)
		Qwen-VL	Regular	84.73 (±0.36)	95.61 (±0.45)	72.81 (±0.38)	82.67 (±0.41)
			VCD	**88.63** (±0.10)	94.64 (±0.25)	81.91 (±0.19)	**87.81** (±0.11)
		InstructBLIP	Regular	80.71 (±0.73)	81.67 (±0.67)	79.19 (±1.14)	80.41 (±0.80)
			VCD	**84.53** (±0.38)	88.55 (±0.54)	79.32 (±0.44)	**83.68** (±0.40)
	Popular	LLaVA1.5	Regular	81.88 (±0.48)	88.93 (±0.60)	72.80 (±0.57)	80.06 (±0.05)
			VCD	**85.38** (±0.38)	86.92 (±0.53)	83.28 (±0.42)	**85.06** (±0.37)
		Qwen-VL	Regular	84.13 (±0.18)	94.31 (±0.43)	72.64 (±0.45)	82.06 (±0.23)
			VCD	**87.12** (±0.07)	91.49 (±0.10)	81.85 (±0.19)	**86.40** (±0.09)
		InstructBLIP	Regular	78.22 (±0.84)	77.87 (±1.03)	78.85 (±0.52)	78.36 (±0.76)
			VCD	**81.47** (±0.42)	82.89 (±0.64)	79.32 (±0.44)	**81.07** (±0.39)
	Adversarial	LLaVA1.5	Regular	78.96 (±0.52)	83.06 (±0.58)	72.75 (±0.59)	77.57 (±0.57)
			VCD	**80.88** (±0.33)	79.45 (±0.29)	83.29 (±0.43)	**81.33** (±0.34)
		Qwen-VL	Regular	82.26 (±0.30)	89.97 (±0.33)	72.61 (±0.50)	80.37 (±0.37)
			VCD	**84.26** (±0.39)	85.84 (±0.45)	82.05 (±0.39)	**83.90** (±0.39)
		InstructBLIP	Regular	75.84 (±0.45)	74.30 (±0.63)	79.03 (±0.68)	76.59 (±0.40)
			VCD	**79.56** (±0.41)	79.67 (±0.59)	79.39 (±0.50)	**79.52** (±0.38)

Table from: https://arxiv.org/pdf/2311.16922

What is MME

- MME is a benchmark for assessing LVLMs in both perception (e.g., object existence, counting, attributes) and cognition tasks.

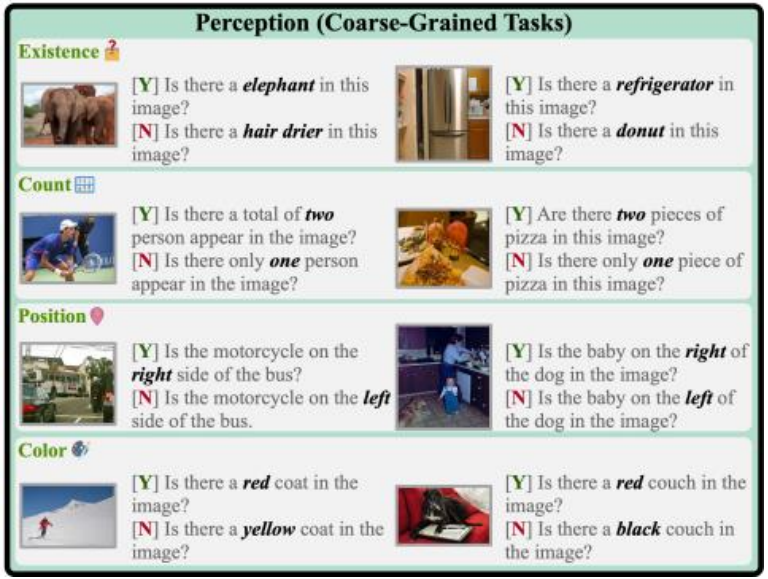


Figure from: https://arxiv.org/pdf/2311.16922

MME Results

- VCD outperforms regular decoding on object- and attribute-level tasks.
- LLaVA-1.5's score jumps from ~565 to ~605.
- Strong color accuracy gains.

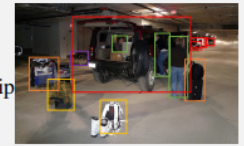
Model	Decoding	Object-level		Attribute-level		Total Scores↑
		Existence↑	*Count*↑	*Position*↑	*Color*↑	
LLaVA1.5	Regular	$175.67_{(\pm 7.51)}$	$124.67_{(\pm 19.59)}$	$114.00_{(\pm 9.32)}$	$151.00_{(\pm 10.45)}$	$565.33_{(\pm 32.92)}$
	VCD	$\mathbf{184.66}_{(\pm 6.81)}$	$\mathbf{138.33}_{(\pm 15.68)}$	$\mathbf{128.67}_{(\pm 7.21)}$	$\mathbf{153.00}_{(\pm 7.58)}$	$\mathbf{604.66}_{(\pm 18.76)}$
Qwen-VL	Regular	$155.00_{(\pm 3.54)}$	$127.67_{(\pm 13.36)}$	$\mathbf{131.67}_{(\pm 7.73)}$	$173.00_{(\pm 9.75)}$	$587.33_{(\pm 31.06)}$
	VCD	$\mathbf{156.00}_{(\pm 6.52)}$	$\mathbf{131.00}_{(\pm 6.19)}$	$128.00_{(\pm 3.61)}$	$\mathbf{181.67}_{(\pm 5.14)}$	$\mathbf{596.67}_{(\pm 11.61)}$
InstructBLIP	Regular	$141.00_{(\pm 13.97)}$	$75.33_{(\pm 14.16)}$	$\mathbf{66.67}_{(\pm 3.91)}$	$97.33_{(\pm 16.94)}$	$380.33_{(\pm 40.20)}$
	VCD	$\mathbf{168.33}_{(\pm 11.55)}$	$\mathbf{92.33}_{(\pm 8.47)}$	$64.00_{(\pm 6.73)}$	$\mathbf{123.00}_{(\pm 11.27)}$	$\mathbf{447.67}_{(\pm 13.36)}$

Table from: https://arxiv.org/pdf/2311.16922

What is LLaVA-Bench

- LLaVA-Bench is a small, diverse set (24 images, 60 questions) for testing open-ended vision–language capabilities.

Context type 1: Captions
A group of people standing outside of a black vehicle with various luggage.
Luggage surrounds a vehicle in an underground parking area
People try to fit all of their luggage in an SUV.
The sport utility vehicle is parked in the public garage, being packed for a trip
Some people with luggage near a van that is transporting it.



Context type 2: Boxes
person: [0.681, 0.242, 0.774, 0.694], backpack: [0.384, 0.696, 0.485, 0.914], suitcase: ...<omitted>

Response type 1: conversation
Question: What type of vehicle is featured in the image?
Answer: The image features a black sport utility vehicle (SUV) ...<omitted>
Response type 2: detailed description
The image is an underground parking area with a black sport utility vehicle (SUV) parked. There are three people in the scene, with one person standing closer to the left side of the vehicle, another person in the middle, and the third person on the right side. They are all working together to pack their luggage into the SUV for a trip. ...<omitted>
Response type 3: complex reasoning
Question: What challenges do these people face?
Answer: In the image, a group of people is standing outside a black SUV in a parking area, surrounded by various pieces of luggage, including suitcases and backpacks. They are facing the challenge of fitting all their luggage into the black SUV. There are multiple suitcases and backpacks to be packed, which suggests that the group has a significant amount of belongings ...<omitted>

Figure from: https://arxiv.org/pdf/2304.08485

LLaVA-Bench

- GPT-4V checks accuracy/detail, and VCD raises scores for LLaVA-1.5, InstructBLIP, Qwen-VL, cutting hallucinations.
- Qwen-VL's accuracy jumps from 4.76 to 6.69 under VCD, showing strong impact.

Model	Decoding	Accuracy↑	Detailedness↑
LLaVA-1.5	Regular	3.23	3.54
	VCD	**4.15**	**3.85**
InstructBLIP	Regular	3.84	4.07
	VCD	**4.23**	**4.69**
Qwen-VL	Regular	4.76	3.46
	VCD	**6.69**	**4.46**

Table from: https://arxiv.org/pdf/2311.16922

Results – Overall Performance

- **Consistent Gains:** VCD cuts hallucinations on POPE, MME, and LLaVA-Bench.
- **No Trade-Off:** It preserves or improves recognition/reasoning.
- **Model-Agnostic:** Works well on LLaVA-1.5, Qwen-VL, InstructBLIP.
- **Stable:** Ablations show robustness to hyperparameters and decoding strategies.
- **Training-Free:** Effectively boosts faithfulness without harming broader abilities.

Ablation Studies
&
Discussion

Ablation Studies – Noise Steps (T)

- Higher T increase visual uncertainty.
- Too much distortion harms performance.

T	*Hallucination Subset Total Scores*	*Perception Subset Total Scores*	*Recognition Subset Total Scores*
200	$586.67_{\pm 11.67}$	$1311.47_{\pm 4.33}$	$338.69_{\pm 19.87}$
500	$591.67_{\pm 36.06}$	$1340.89_{\pm 55.91}$	$323.45_{\pm 5.89}$
700	$578.89_{\pm 19.17}$	$1339.04_{\pm 40.40}$	$320.95_{\pm 14.18}$
999	$557.78_{\pm 1.92}$	$1345.81_{\pm 36.31}$	$321.90_{\pm 10.19}$

Table from: https://arxiv.org/pdf/2311.16922

Ablation Studies – Effect of α

- **Moderate α** give the **best trade-off** between correctness and coherence.

$$p_{vcd}(y \mid v, v', x) = softmax[(1+\alpha)\, logit_\theta(y \mid v, x) - \alpha\, logit_\theta(y \mid v', x)]$$

α	*Hallucination Subset Total Scores*	*Perception Subset Total Scores*	*Recognition Subset Total Scores*
0.25	$583.89_{\pm 19.32}$	$1322.25_{\pm 32.58}$	$330.24_{\pm 13.60}$
0.5	$580.56_{\pm 17.11}$	$1315.49_{\pm 27.28}$	$333.45_{\pm 5.77}$
0.75	$578.33_{\pm 29.49}$	$1312.93_{\pm 39.31}$	$330.95_{\pm 13.58}$
1.0	$591.67_{\pm 36.06}$	$1340.89_{\pm 55.91}$	$323.45_{\pm 5.89}$

Table & Equation from: https://arxiv.org/pdf/2311.16922

Ablation Studies – Effect of β

- β filters **out low-confidence** tokens to protect valid content from over-penalization.
- **Moderate β** give the **best trade-off** between correctness and coherence.

$$V_{head}(y_{<t}) = \{y_t \in V : p_\theta(y_t \mid v, x, y_{<t}) \geq \beta \max_w p_\theta(w \mid v, x, y_{<t})\}$$

β	*Hallucination Subset Total Scores*	*Perception Subset Total Scores*	*Recognition Subset Total Scores*
0	$577.22_{\pm 11.10}$	$1299.04_{\pm 39.30}$	$302.98_{\pm 19.82}$
0.001	$574.44_{\pm 6.31}$	$1298.71_{\pm 40.24}$	$289.88_{\pm 14.02}$
0.01	$583.33_{\pm 18.78}$	$1324.44_{\pm 37.84}$	$327.38_{\pm 17.11}$
0.1	$591.67_{\pm 36.06}$	$1340.89_{\pm 55.91}$	$323.45_{\pm 5.89}$
0.2	$591.67_{\pm 7.26}$	$1343.06_{\pm 13.06}$	$328.57_{\pm 16.37}$
0.5	$635.00_{\pm 7.64}$	$1474.02_{\pm 15.53}$	$331.43_{\pm 13.03}$

Table & Equation from: https://arxiv.org/pdf/2311.16922

Ablation Studies – Sampling Strategies

- VCD **maintains** or boosts **performance** under diverse sampling strategies.
- **Stable** gains persist even with temperature adjustments, reflecting robust hallucination reduction.

Sampling Strategy	w. VCD	Accuracy	Precision	Recall	F1 Score
Top P	No	$84.91_{\pm 0.25}$	$94.73_{\pm 0.30}$	$73.93_{\pm 0.52}$	$83.05_{\pm 0.32}$
	Yes	$\mathbf{87.82}_{\pm 0.66}$	$91.17_{\pm 0.57}$	$83.76_{\pm 0.87}$	$\mathbf{87.31}_{\pm 0.72}$
Top K	No	$83.04_{\pm 0.16}$	$91.84_{\pm 0.15}$	$72.53_{\pm 0.44}$	$81.05_{\pm 0.24}$
	Yes	$\mathbf{87.49}_{\pm 0.56}$	$91.09_{\pm 0.53}$	$83.11_{\pm 0.71}$	$\mathbf{86.92}_{\pm 0.60}$
Greedy	No	$87.10_{\pm 0.00}$	$97.33_{\pm 0.00}$	$76.29_{\pm 0.00}$	$85.54_{\pm 0.00}$
	Yes	$\mathbf{88.49}_{\pm 0.28}$	$91.78_{\pm 0.28}$	$84.56_{\pm 0.44}$	$\mathbf{88.02}_{\pm 0.30}$
Top K+Temperature 0.7	No	$85.17_{\pm 0.12}$	$94.82_{\pm 0.12}$	$74.40_{\pm 0.35}$	$83.38_{\pm 0.17}$
	Yes	$\mathbf{87.94}_{\pm 0.51}$	$91.21_{\pm 0.49}$	$83.98_{\pm 0.60}$	$\mathbf{87.45}_{\pm 0.54}$
Top K+Temperature 1.5	No	$79.28_{\pm 0.22}$	$86.48_{\pm 1.12}$	$69.42_{\pm 0.91}$	$77.01_{\pm 0.22}$
	Yes	$\mathbf{86.97}_{\pm 0.50}$	$90.96_{\pm 0.64}$	$82.09_{\pm 0.41}$	$\mathbf{86.30}_{\pm 0.51}$

Table from: https://arxiv.org/pdf/2311.16922

Ablation Studies – Model Scale

- **Scaling up** models still benefits from VCD, showing no drop in effectiveness.

Dataset	POPE	Model	Decoding	Accuracy	Precision	Recall	F1 Score
MSCOCO	*Random*	LLaVA1.5(13B)	Regular	$83.31_{\pm 0.32}$	$91.46_{\pm 0.38}$	$73.48_{\pm 0.75}$	$81.49_{\pm 0.43}$
			VCD	$\mathbf{87.39}_{\pm 0.32}$	$92.68_{\pm 0.36}$	$81.19_{\pm 0.63}$	$\mathbf{86.55}_{\pm 0.41}$
		InstructBLIP(13B)	Regular	$82.36_{\pm 0.59}$	$86.93_{\pm 0.85}$	$76.19_{\pm 1.05}$	$81.20_{\pm 0.68}$
			VCD	$\mathbf{84.53}_{\pm 0.38}$	$88.55_{\pm 0.54}$	$79.32_{\pm 0.44}$	$\mathbf{83.68}_{\pm 0.40}$
	Popular	LLaVA1.5(13B)	Regular	$82.47_{\pm 0.55}$	$89.55_{\pm 0.92}$	$73.53_{\pm 0.78}$	$80.75_{\pm 0.61}$
			VCD	$\mathbf{85.74}_{\pm 0.25}$	$89.33_{\pm 0.52}$	$81.19_{\pm 0.63}$	$\mathbf{85.06}_{\pm 0.29}$
		InstructBLIP(13B)	Regular	$79.07_{\pm 0.66}$	$81.11_{\pm 0.70}$	$75.79_{\pm 1.27}$	$78.35_{\pm 0.78}$
			VCD	$\mathbf{81.47}_{\pm 0.42}$	$82.89_{\pm 0.64}$	$79.32_{\pm 0.44}$	$\mathbf{81.07}_{\pm 0.39}$
	Adversarial	LLaVA1.5(13B)	Regular	$80.00_{\pm 0.52}$	$84.46_{\pm 0.73}$	$73.53_{\pm 0.76}$	$78.62_{\pm 0.58}$
			VCD	$\mathbf{81.92}_{\pm 0.44}$	$82.40_{\pm 0.42}$	$81.17_{\pm 0.65}$	$\mathbf{81.78}_{\pm 0.47}$
		InstructBLIP(13B)	Regular	$76.57_{\pm 0.75}$	$77.00_{\pm 0.83}$	$75.79_{\pm 0.80}$	$76.39_{\pm 0.75}$
			VCD	$\mathbf{79.56}_{\pm 0.41}$	$79.67_{\pm 0.59}$	$79.39_{\pm 0.50}$	$\mathbf{79.52}_{\pm 0.38}$

Table from: https://arxiv.org/pdf/2311.16922

Future Directions

- **Advanced Distortion Techniques:** Investigate object-level or region-wise perturbations
- **Video & Multimedia Expansion:** Adapt VCD to temporal domains (e.g., videos)
- **Hybrid Approaches:** Combine VCD with retrieval or RLHF for enriched factual grounding
- **Cross-Modality & Cross-Lingual:** Explore more diverse modalities and languages

Limitations
&
Conclusion

Limitations

- Currently **relies** on **Gaussian noise** for distortion, **leaving** room for more **nuanced** methods.
- Primarily **tested** on **single-image** tasks; **extending** to **video** or other modalities remains unexplored.
- **Requires** strong **visual encoding**; **performance** may **vary** with different **vision backbones**.
- Further **analysis** needed **to refine** adaptive **constraints** and handle **subtle visual** cues.

Conclusion

- **VCD yields +7.4 F1** improvement on **POPE** and **+18% on MME**, demonstrating robust performance.
- **VCD** is **training-free** and **lightweight**, eliminating the need for additional datasets or models.
- **VCD generalizes** across **LLaVA-1.5**, **InstructBLIP**, and **Qwen-VL**, highlighting its wide applicability.
- **VCD** also simultaneously **improves** general **perception tasks** (e.g., **object existence**, **color**).

Q & A

Thank you for listening!