Multi-Object Hallucination in Vision-Language Models

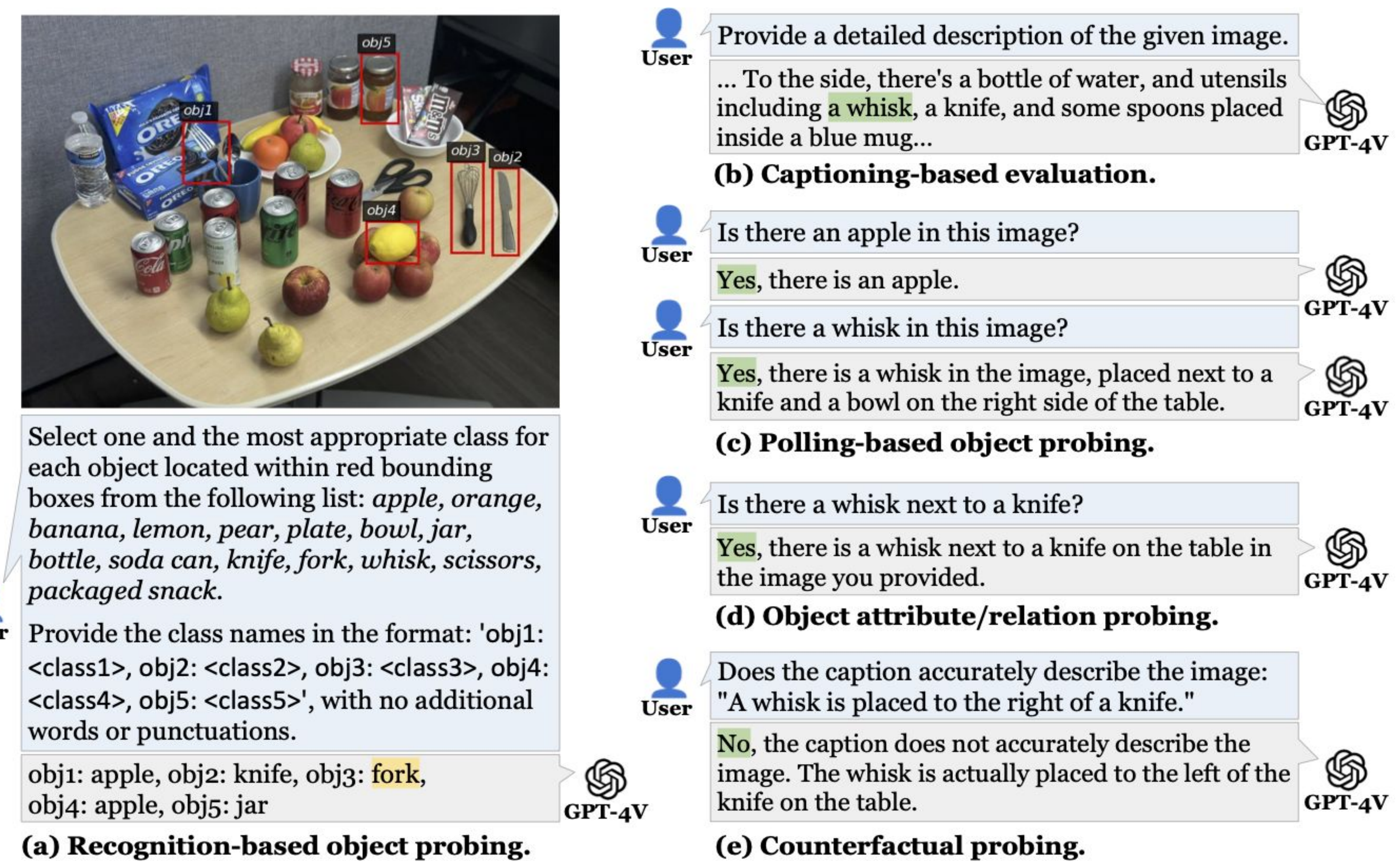
Xuweiyi Chen*[1,2], Ziqiao Ma*[1], Xuejun Zhang*[1] (*equal contributions)
Sihan Xu[1], Shengyi Qian[1,3], Jianing Yang[1], David Fouhey[3], Joyce Chai[1]

1 University of Michigan 2 University of Virginia 3 NYU

data & code

NEURAL INFORMATION PROCESSING SYSTEMS

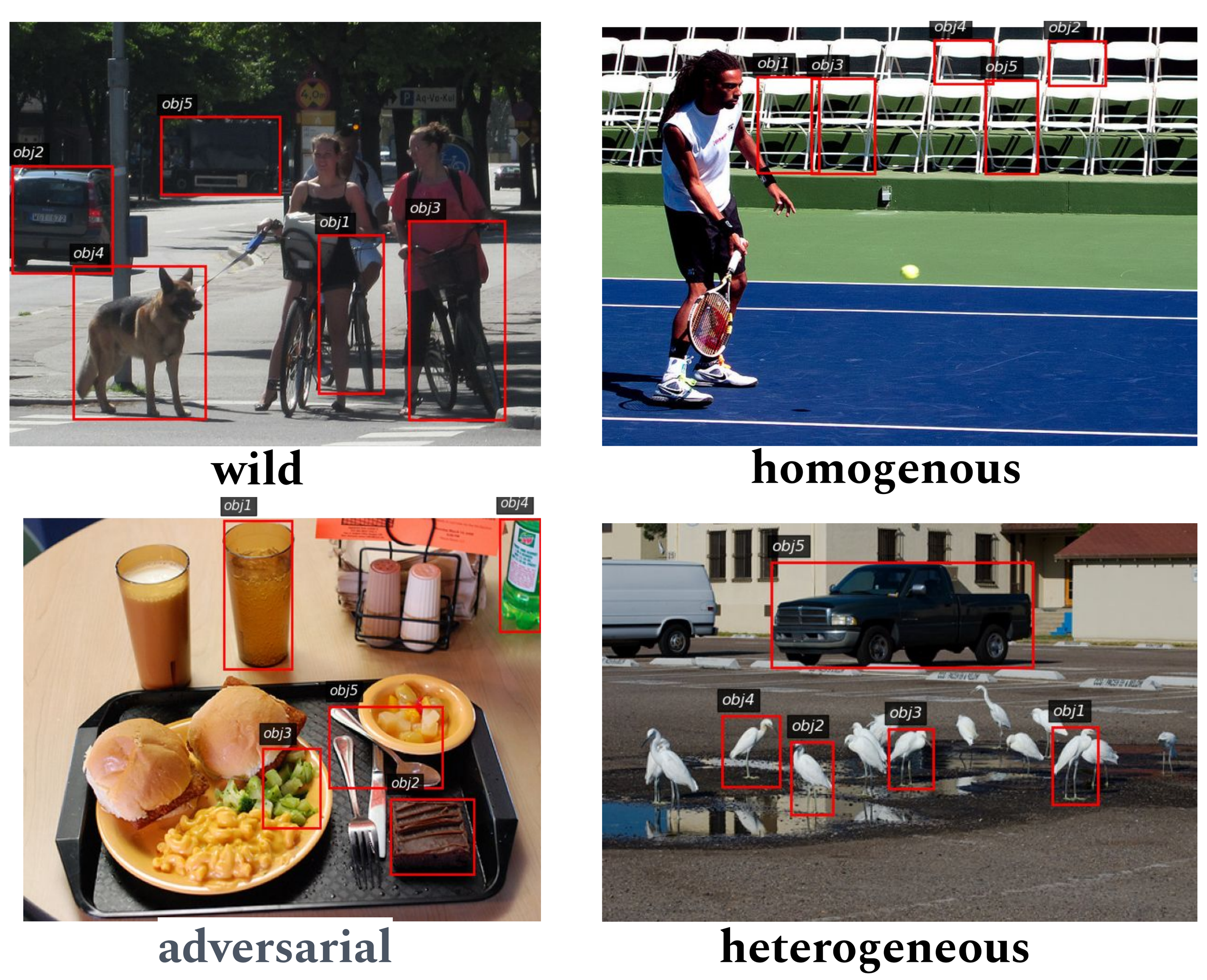
Motivation



- Grounding is not simply one-to-one between objects and classes, but a many-to-many mapping between objects and phrases.
- LVLMs suffer more hallucinations in multi-object task than in single-object ones.

ROPE: Recognition-based Object Probing Evaluation Benchmark

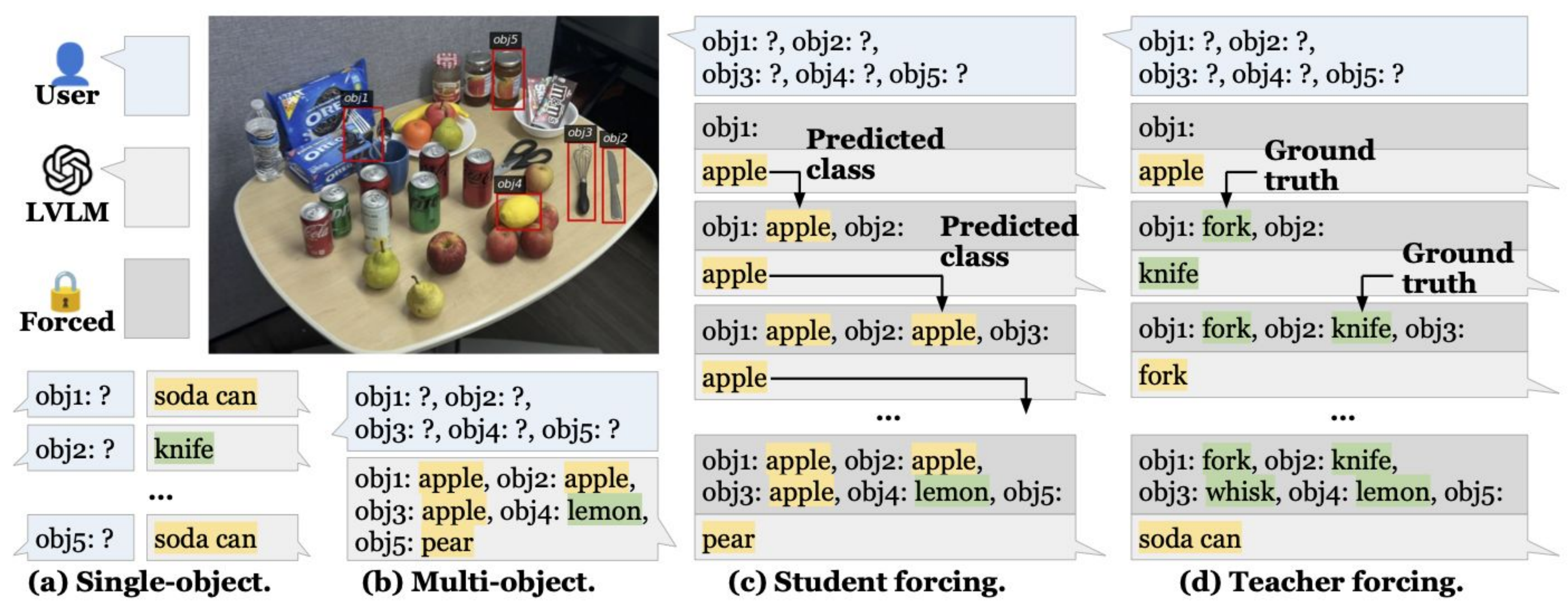
A dataset designed to evaluate multi-object hallucination in LVLMs, challenging models to recognize objects in homogeneous, heterogeneous, and adversarial scenarios.



wild homogenous adversarial heterogeneous

Experiment Setting

Different types of instruction settings of ROPE.



(a) Single-object. (b) Multi-object. (c) Student forcing. (d) Teacher forcing.

- Single-object Probing: LVLMs are tasked to recognize one object at a time.
- Multi-object Probing: LVLMs must simultaneously recognize multiple objects in a single prompt.
- Student Forcing: LVLMs are instructed to decode only the object tokens in a forced format without output manipulation.
- Teacher Forcing: LVLMs generate object tokens while being conditioned on the correct previous context.

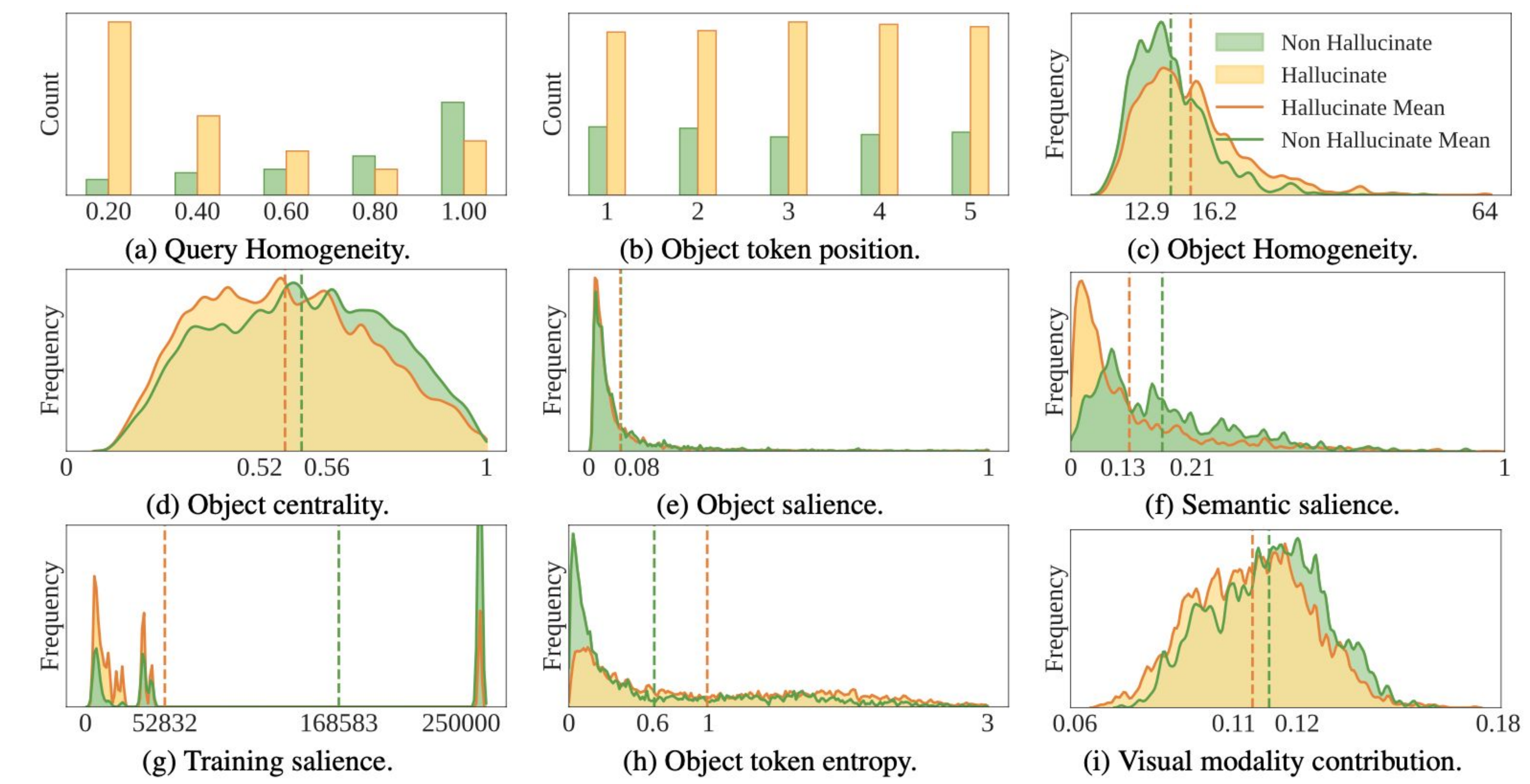
Experiment Result

The experiments evaluate LVLM performance on seen and unseen data across varying object distributions, such as Wild, Homogeneous, and Heterogeneous scenarios.

Models	Default Multi-Object			Student-Forcing			Teacher-Forcing			Single-Object		
	Wild	Hom.	Het.	Wild	Hom.	Het.	Wild	Hom.	Het.	Wild	Hom.	Het.
Seen												
Yi-VL-6B	2.95	5.65	1.99	3.44	6.80	3.78	5.45	26.25	4.36	0.19	0.30	0.13
Yi-VL-34B	8.50	15.35	3.33	8.97	16.30	4.23	10.09	19.75	4.94	0.22	2.60	0.13
LLaVA-7B	31.29	67.50	8.00	31.28	67.25	11.22	31.49	92.15	12.37	35.32	62.35	17.37
LLaVA-13B	31.54	67.63	12.64	31.49	73.25	11.54	34.97	94.25	16.03	43.13	80.60	23.91
LLaVA-34B	39.95	85.75	18.85	**52.75**	**85.20**	**33.91**	**56.41**	**95.81**	**25.31**	55.05	**86.50**	18.97
Qwen VL	2.73	6.60	1.03	6.25	16.00	3.65	18.74	71.50	5.45	8.73	16.05	5.58
Qwen VL-C	8.72	16.90	6.67	5.26	8.60	4.10	12.11	47.75	8.08	25.99	43.40	13.21
CogVLM	0.04	0.00	0.00	0.00	0.00	0.00	0.10	0.95	0.00	0.00	0.00	0.00
CogVLM-G	0.00	0.00	0.00	9.86	13.50	6.79	22.64	75.45	0.45	11.25	22.65	7.12
CogVLM-C	12.89	22.75	7.18	25.37	43.63	12.03	28.25	72.80	17.50	30.16	56.00	16.35
LLaVA-7B*	N/A			9.16	16.40	5.51	N/A			11.68	23.55	9.36
GLaMM*	N/A			27.11	53.35	13.01	N/A			**63.81**	81.75	53.40
GroundHOG*	N/A			23.57	30.80	24.23	N/A			44.80	43.10	38.97
IDEFICS	0.00	1.45	0.13	6.25	18.70	0.64	17.37	76.15	10.06	4.62	0.00	0.32
CogVLM-2	21.51	37.55	17.31	37.02	70.85	12.69	37.10	73.50	17.44	21.16	38.75	13.65
MiniCPM-V	34.75	59.91	17.37	31.62	62.80	13.65	32.16	68.05	16.79	27.42	55.35	16.92
GPT4V**	53.80	77.55	40.83	N/A			N/A			55.89	78.25	41.03
GPT4O**	**71.27**	**89.25**	**66.03**	N/A			N/A			60.77	73.92	**54.31**

Hallucinatory Factors Analysis

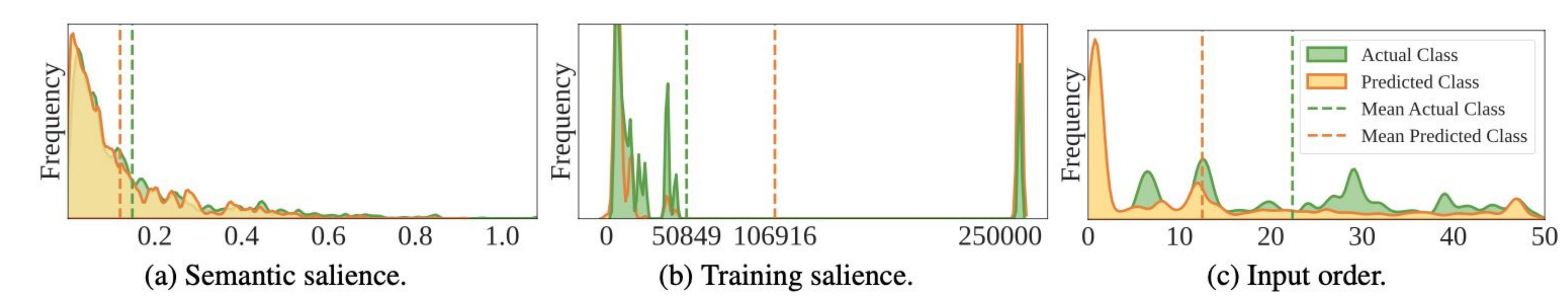
This section analyzes factors contributing to hallucinations in LVLMs, focusing on data/task-specific elements like query homogeneity, salience and frequency of objects, and intrinsic behaviors such as token entropy and visual attention, revealing how these aspects influence hallucination likelihood.



(a) Query Homogeneity. (b) Object token position. (c) Object Homogeneity. (d) Object centrality. (e) Object salience. (f) Semantic salience. (g) Training salience. (h) Object token entropy. (i) Visual modality contribution.

Hallucination Pattern

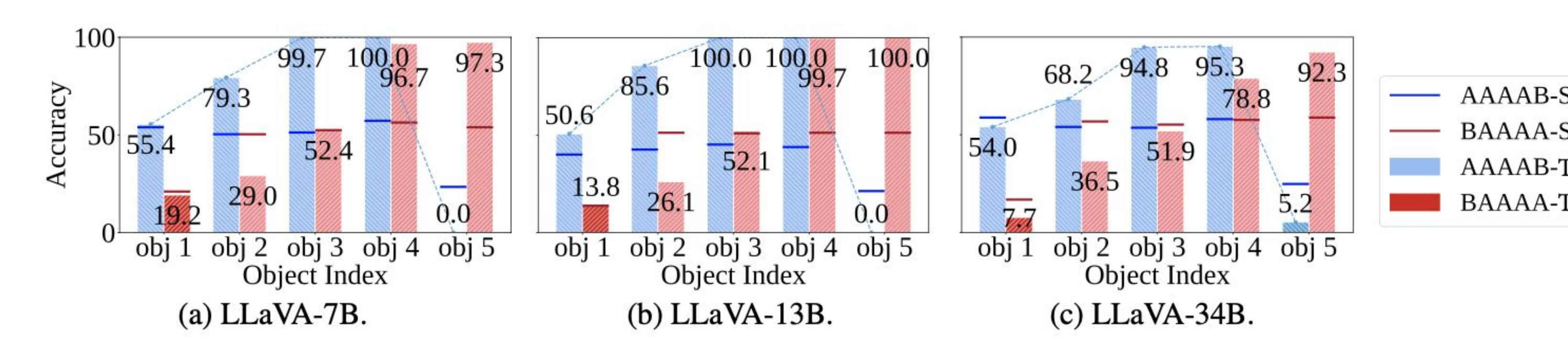
This section examines the distribution of actual versus predicted object classes, highlighting how factors like semantic salience, training salience, and input order contribute to hallucinations in object recognition tasks.



(a) Semantic salience. (b) Training salience. (c) Input order.

Adversarial Performance Analysis

This section illustrates how LVLMs perform under adversarial sequences, where object recognition accuracy drops significantly for the last object in query sequence of AAAAB.



(a) LLaVA-7B. (b) LLaVA-13B. (c) LLaVA-34B.