ODE: Open-Set Evaluation of Hallucinations in Multimodal Large Language Models

Supplementary Material

A. Contamination and Synthetic Validity

Fig. 8 highlights the performance of MiniGPT-4 and InstructBLIP on discriminative tasks across COCO 2014, Internet, and ODE-generated datasets. Internet data, as the most recent and reliably crawled dataset, avoids contamination, making it a more trustworthy baseline compared to COCO 2014, which shows artificially inflated results due to potential overlaps with training data. While ODE-generated synthetic data exhibits slightly lower performance than Internet data, feature space validates the close similarity between synthetic and natural images. This confirms that synthetic data can effectively replicate real-world distributions. Moreover, its controllability and distribution diversity establish it as a valuable resource for evaluating model reliability, particularly in mitigating data contamination and enabling the creation of novel and challenging testing scenarios.

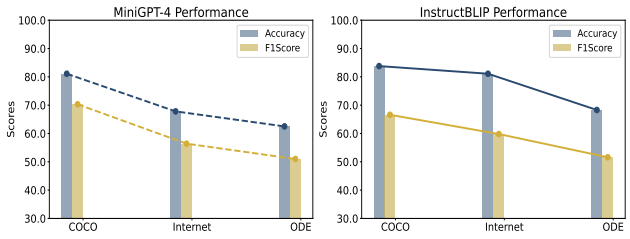

Figure 8. Model performances on discriminative tasks.

The degradation results primarily indicate potential data contamination in certain models. Since the Internet and synthetic images do not share entirely identical semantics, some degree of discrepancy is expected due to differences in specific concepts within the images. We compared the features of three types of images: real images before contamination, contaminated images, and synthetic images generated on the basis of detailed semantic descriptions. The visualization shows that contaminated images form a distinct distribution, while synthetic images align closely with uncontaminated real images. Thus, we conclude that synthetic images are suitable for hallucination evaluation, a perspective also supported by previous works [27].

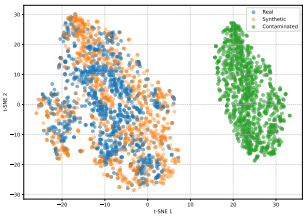

Figure 9. Feature Visualization.

B. Evaluation Results

Table 6 summarizes the detailed evaluation results of five models under the following conditions: Standard, Random, Fictional, and Long-tail distributions. Tables 7, 8, 9, and 10 present the evaluation results of object-level hallucinations on ODE at the attribute level.

From the detailed results, we derive the following key observations:

- High Variability Across Attributes: Models excel in state-related attributes due to strong semantic associations in training data. However, action and number attributes expose significant weaknesses, especially in rare or unseen scenarios.
- Impact of Data Distribution: Standard distributions enable high performance as a result of frequent exposure during training. In contrast, Fictional and Random distributions cause sharp performance declines, exposing over-reliance on memorized correlations. Long-tail distributions further highlight the models' inability to generalize effectively to sparse data.
- Task-Specific Limitations: Generative tasks emphasize semantic fluency, whereas discriminative tasks require detailed attribute recognition. These differences result in varying performance gaps between task types.
- Model-Specific Trends: LLaVA-1.5 demonstrates the best overall performance with balanced precision and recall across attributes and distributions. InstructBLIP achieves high precision but suffers from poor recall, indicating a tendency toward overfitting. MiniGPT-4 and mPLUG-Owl struggle significantly with generalization, particularly in Random and Fictional contexts.
- Broader Implications for Hallucination Mitigation: Addressing relational hallucination requires improving the diversity and balance of training datasets. Attribute-specific fine-tuning and robust data augmentation strategies are essential for better generalization.

Table 6. Evaluation results of five models (CogVLM, LLaVA-1.5, mPLUG_Owl, MiniGPT-4, and InstructBLIP) under the following conditions: Standard, Random, Fictional, and Long-tail distributions. Each task type—Generative Task (CHAIR, Cover, Hal, Cog) and Discriminative Task (Existence and Attribute)—is evaluated with Accuracy (Acc), Precision (P), Recall (R), and F1-score (F1).

		Generative Task				Discriminative-Existence Task				Discriminative-Attribute Task			
Criterion	**Model**	**CHAIR**	**Cover**	**Hal**	**Cog**	**Acc**	**P**	**R**	**F1**	**Acc**	**P**	**R**	**F1**
Standard	CogVLM	51.9	76.5	89.1	12.2	50.7	**100.0**	26.2	41.5	55.2	46.8	55.6	50.8
	LLaVA-1.5	38.9	**77.7**	82.7	8.6	69.5	97.8	55.4	70.7	62.9	32.6	71.9	44.8
	mPLUG	50.8	77.2	96.0	11.5	41.7	94.7	13.4	23.5	**72.5**	41.7	28.6	33.9
	MiniGPT-4	49.4	76.0	93.6	14.2	64.5	97.5	48.0	64.3	69.5	39.7	12.5	19.0
	InstructBLIP	59.9	75.7	88.1	11.0	66.7	96.8	51.7	67.4	60.8	28.5	51.2	36.6
	InternVL-2.5	38.5	60.9	**72.3**	**4.0**	75.7	96.1	66.3	78.4	65.9	48.1	87.0	61.9
	Qwen2-VL	44.2	75.2	88.6	11.6	75.2	97.0	64.9	77.7	67.0	**49.6**	86.8	**63.1**
	Cambrian	**36.3**	64.9	79.2	7.3	**76.9**	96.5	**67.8**	**79.6**	65.7	47.7	**90.4**	62.4
Random	CogVLM	58.1	57.7	87.6	6.0	40.0	89.7	18.0	30.0	57.1	82.5	38.2	52.2
	LLaVA-1.5	45.2	57.7	84.2	4.7	74.7	89.7	69.6	78.3	**75.7**	90.1	65.9	**76.1**
	mPLUG	57.9	56.4	92.1	6.3	40.0	84.0	10.8	19.1	48.4	76.5	20.2	31.9
	MiniGPT-4	50.3	57.9	82.2	5.8	66.4	86.9	58.2	69.7	45.0	74.5	10.8	18.8
	InstructBLIP	55.9	58.4	83.2	5.6	64.6	87.6	54.6	67.2	72.3	**91.5**	58.6	71.4
	InternVL-2.5	43.0	58.4	67.8	3.7	81.4	**89.9**	80.7	85.0	72.8	67.0	80.6	73.1
	Qwen2-VL	38.5	**60.9**	72.3	4.0	**82.8**	88.8	**83.5**	**86.0**	71.4	65.8	**82.6**	73.2
	Cambrian	**30.9**	56.9	**61.4**	**2.3**	80.7	87.0	83.0	84.9	71.2	63.4	84.3	72.3
Fictional	CogVLM	54.0	55.9	80.7	6.3	39.5	90.9	16.1	27.3	50.5	75.4	28.7	41.5
	LLaVA-1.5	48.0	54.2	79.2	4.8	72.5	87.9	66.7	75.8	**73.0**	88.5	62.0	**72.9**
	mPLUG	59.2	54.0	90.6	6.6	42.5	**95.7**	11.8	21.0	45.8	72.1	17.4	28.0
	MiniGPT-4	48.0	55.4	79.2	7.3	67.1	87.4	58.1	69.8	42.9	75.0	9.7	17.2
	InstructBLIP	59.7	56.9	80.2	5.9	66.0	88.6	56.2	68.7	71.0	**88.8**	58.4	70.4
	InternVL-2.5	43.7	55.4	65.3	4.5	78.6	86.5	79.3	82.7	66.2	58.0	73.7	64.9
	Qwen2-VL	41.5	**60.4**	69.8	5.3	77.2	86.6	76.6	81.2	65.8	59.0	75.0	66.0
	Cambrian	**34.5**	55.0	**63.4**	**3.4**	**79.1**	86.6	**80.1**	**83.2**	67.1	58.3	**77.2**	66.4
Long-tail	CogVLM	54.8	67.1	89.6	14.3	41.4	90.0	15.6	26.6	61.5	77.3	49.0	59.9
	LLaVA-1.5	**44.5**	71.3	91.1	11.2	51.3	84.7	32.9	47.4	**78.2**	**85.5**	74.3	**79.5**
	mPLUG	51.9	71.0	96.0	10.7	38.1	91.4	7.9	14.5	48.9	67.4	24.8	36.2
	MiniGPT-4	48.1	**75.2**	93.1	16.5	63.4	**92.9**	48.8	63.9	48.0	75.9	17.1	27.9
	InstructBLIP	51.3	71.0	87.1	11.2	51.5	91.1	30.4	45.6	73.3	84.6	65.1	73.5
	InternVL-2.5	50.4	70.3	86.1	11.1	**70.1**	88.1	63.9	74	65.2	50.5	**89.7**	64.6
	Qwen2-VL	49.8	72.3	93.1	16.7	69.3	92.2	58.9	**71.8**	61.4	48.2	82.9	60.9
	Cambrian	45.0	53.0	**80.7**	**9.0**	68.5	89.6	**59.7**	71.6	60.8	47.0	88.7	61.4

Table 7. The detailed evaluation results of ODE on the object hallucination (attribute-level) under Standard distribution.

	Metric	mPLUG-Owl	MiniGPT-4	LLaVA-1.5	CogVLM	InstructBLIP
State	Acc	57.8	51.3	**80.5**	69.5	71.3
	P	75.3	84.2	86.8	**87.0**	86.8
	R	28.6	12.5	**71.9**	55.6	51.2
	F1	41.4	21.7	**78.6**	67.8	64.4
Number	Acc	40.4	39.9	**92.6**	54.5	87.0
	P	82.7	93.5	80.2	98.4	**98.5**
	R	15.3	10.6	**90.3**	32.7	81.9
	F1	25.8	19.0	**94.1**	49.1	89.4
Action	Acc	53.3	43.9	**66.9**	60.1	67.3
	P	67.8	45.8	**74.2**	66.9	74.5
	R	17.5	9.6	**51.8**	51.3	52.6
	F1	27.8	15.8	61.0	58.0	**61.6**

Table 8. The detailed evaluation results of ODE on the object hallucination (attribute-level) under Long-tail distribution.

	Metric	mPLUG-Owl	MiniGPT-4	LLaVA-1.5	CogVLM	InstructBLIP
State	Acc	55.7	50.5	**75.5**	66.1	64.2
	P	69.1	71.4	82.5	**84.3**	74.6
	R	32.2	13.7	**64.7**	53.4	43.2
	F1	43.9	23.0	**72.5**	65.3	54.7
Number	Acc	40.1	42.4	**86.8**	49.8	83.7
	P	78.9	85.7	95.5	91	**97.2**
	R	14.9	16.3	**94.2**	27.5	78
	F1	25	27.44	**89.4**	42.2	86.5
Action	Acc	48.5	51.9	**64.4**	61	**64.4**
	P	50.7	62.8	64.2	61.8	**64.6**
	R	27.3	20.5	65.2	**71.2**	63.6
	F1	35.4	30.9	64.6	**66.1**	64.0

Table 9. The detailed evaluation results of ODE on the object hallucination (attribute-level) under Fictional distribution.

	Metric	mPLUG-Owl	MiniGPT-4	LLaVA-1.5	CogVLM	InstructBLIP
State	Acc	51.3	50.2	**71.1**	55.0	66.4
	P	67.5	76.5	**87.7**	84.4	85.7
	R	22.4	11.2	**49.1**	28.0	41.4
	F1	33.6	19.5	**62.9**	42	55.8
Number	Acc	39.9	35.8	**76.9**	46	73.4
	P	85.7	72.7	91.5	86.6	**93.6**
	R	13.4	5.9	**72.0**	24.0	65.1
	F1	23.2	10.9	**80.5**	37.6	76.7
Action	Acc	45.7	46.7	**56.5**	42.4	57.6
	P	30.0	**90.9**	60. 0	46.5	62.5
	R	6.5	21.7	39.1	**43.5**	**43.5**
	F1	10.7	35.0	47.3	44.9	**51.2**

Table 10. The detailed evaluation results of ODE on the object hallucination (attribute-level) under Random distribution.

	Metric	mPLUG-Owl	MiniGPT-4	LLaVA-1.5	CogVLM	InstructBLIP
State	Acc	55.0	53.5	**71.9**	60.3	66.1
	P	76.4	78.3	86.3	85.0	**87.5**
	R	22.7	14.9	**52.1**	39.7	40.5
	F1	35.0	25.0	**64.9**	54.1	55.3
Action	Acc	40.9	36.8	**80.4**	48.0	75.2
	P	79.8	78.4	92.5	87.2	**95.0**
	R	16.6	7.2	**76.7**	27.0	66.8
	F1	27.5	13.2	**83.8**	41.2	78.4
Number	Acc	54.7	45.3	54.7	65.6	**60.9**
	P	75	0	75	**76.9**	**76.9**
	R	18.7	0	18.7	**62.5**	31.2
	F1	29.9	0	29.9	**68.9**	44.3