PAS : Prelim Attention Score for Detecting Object Hallucinations in Large Vision–Language Models

Supplementary Material

Contents

A. Extended discussion of Related Work

This section provides a more comprehensive discussion of related work to supplement the main paper. In particular, we elaborate on some hallucination *mitigation* works that made some observations related to the prelim but did not reach our conclusions. This discussion is intended to clarify the precise novelty of our work and provide a more complete picture of the surrounding literature.

"Text inertia." Coined by PAI [25], this term refers to the phenomenon where the LVLM produces the same hallucination as its backbone LLM. They characterize this phenomenon by a lack of attention to the image, and propose manually increasing the image attention to mitigate hallucinations. They use the "pure" LLM (before vision adaptation) to define what would be the "prediction" of the LVLM without the image.

Our work expands on PAI in two substantial ways. First, we directly look at overdependence on prelim tokens (**H1**). Our novel MI-based formulation characterizes the relationship between the generated tokens, the image, and the prelim. Second, we do not rely on the notion of an underlying LLM. Thus, our formulation is still meaningful even if the LVLM ignores the image and defaults to something other than its underlying LLM. This is important as many LVLM methods also finetune the underlying LLMs, which potentially makes it behaves differently from the pure LLM.

"Summary tokens." A phenomenon discovered by OPERA [15], in which a particular attention pattern involving some special output tokens (*e.g.*, periods) can indicate hallucination. They propose a decoding scheme that avoids this pattern to mitigate hallucination.

Our work focus on a broader problem, which is the overall role of prelim tokens in (object) hallucination. In context of our work, the "summary tokens" can be viewed as a more specific sub-phenomenon. Furthermore, OPERA's observation focus on sentence-level hallucination, and cannot indicate with token is hallucinatory.

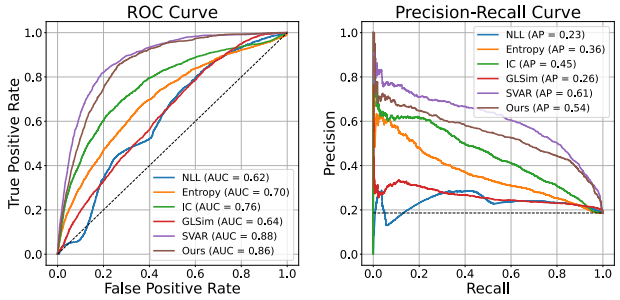

Figure 7. The ROC and PRC curves for object hallucination detection of our method and the baselines for MiniGPT-4 on MSCOCO dataset. Dashed line indicates chance performance.

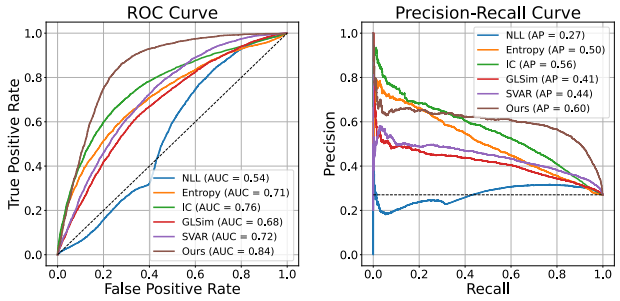

Figure 8. The ROC and PRC curves for object hallucination detection of our method and the baselines for Shikra on MSCOCO dataset. Dashed line indicates chance performance.

B. Additional results

B.1. Results on 13B models

In Table 6 we provide additional comparative evaluation of our method and the baselines on 13B models, with the same setting as Tab. 2. This includes more recent model Cambrian-1 (NeurIPS 2024). The result shows that our method is robust across model sizes and architectures.

B.2. Statistical tests

Significance of scores. In Table 7 and Table 8 we report the Mann-Whitney U statistics, which test the null hypothesis that the score for a real and hallucinated object is the same, for both PAS and MI-based score. This result support

Method	LLaVA-1.5-13B		InstructBLIP 13B		Cambrian-1-13B		Average
	MSCOCO	Pascal VOC	MSCOCO	Pascal VOC	MSCOCO	Pascal VOC	
NLL [37]	56.4	64.5	62.0	66.4	67.5	75.1	65.3
Entropy [26]	71.1	62.8	77.2	74.6	52.4	45.5	63.9
IC [16]	73.2	67.2	73.6	64.7	65.5	61.4	67.6
GLSim [27]	74.5	81.2	84.2	83.2	78.5	**81.6**	80.5
SVAR [17]	81.9	83.6	84.8	83.7	74.6	74.3	80.5
Ours	**83.7**	**86.1**	**86.1**	**85.9**	78.2	79.9	**83.3**

Table 6. Object hallucination detection performance (AUROC, higher is better) for 13B models. All values are percentages, and best results are shown in **bold**.

Model	U statistic	p-value
LLaVA-1.5-7B	42437327.5	< 0.001
Shikra	46474015.5	< 0.001
MiniGPT-4	19288066.0	< 0.001

Table 7. Mann–Whitney U test statistics for PAS, supplementing Fig. 3a.

Model	U statistic	p-value
LLaVA-1.5-7B	38585092.5	< 0.001
Shikra	43494912.0	< 0.001
MiniGPT-4	19845644.5	< 0.001

Table 8. Mann–Whitney U test statistics for MI-based score (Eq. (9)), supplementing Fig. 3b.

our claims in Sec. 4. The statistics and p-values complement Fig. 3, which visually depicts the same distributions.

Significance of improvements. In Table 11 and Table 12 we show the 95% confidence interval (CI) for the improvement in AUROC ($\text{AUROC}_{\text{ours}} - \text{AUROC}_{\text{baseline}}$), computed from the DeLong test. The test's implementation is from the MLstatkit library. The result shows that for the vast majority of cases, our method yields a statistically significant improvement of AUROC over the baselines.

B.3. Additional ROC & PRC curves

We provide additional depictions of the ROC and PRC curves for MiniGPT-4 (Figure 7) and Shikra (Figure 8). Along with Fig. 4, they describe all models in Tab. 2 in the main text. Note that the chance performance (assigning "halluication" to all objects) for the Precision-Recall Curve is determined by the ratio of hallucinated objects to all objects, which is described in Tab. 9 and Tab. 10.

C. Experiment/implementation details

All experiments can be performed on a single GPU with 80GB VRAM. Unless otherwise specified, we use LVLMs with `float16` precision to save memory.

Object counts. We report the number of unique objects (belonging to MSCOCO's 80 object classes) generated for each model across each dataset.

Model	No. of Real	No. of Hallucinated
LLaVA-1.5-7B	12,576	4,008
Shikra	12,184	4,515
MiniGPT-4	9,926	2,269

Table 9. Object hallucination statistics on MSCOCO (5,000 samples).

Model	No. of Real	No. of Hallucinated
LLaVA-1.5-7B	8,376	3,636
Shikra	8,286	3,802
MiniGPT-4	7,680	2,534

Table 10. Object hallucination statistics on Pascal VOC (5,823 samples).

Multi-token objects. LVLMs operate on tokens, while objects can potentially span multiple tokens. However, the first token is arguably most important since the next tokens (if any) need only "continue" what is already generated. Furthermore, most of the 80 MSCOCO object classes have distinct first tokens. Therefore, following [27], we default to using the first token for all methods, including ours and the baselines.

Token types in input. We further clarify the illustration of different token types in Figure 1. First, instruction tokens are defined as input tokens that are not the BOS and image tokens. This means that it includes the prompt ("Please help me describe the image in detail.") and possibly a system prompt, if the LVLM has one.

Method	LLaVA-1.5-7B		MiniGPT-4		Shikra	
	MSCOCO	Pascal VOC	MSCOCO	Pascal VOC	MSCOCO	Pascal VOC
NLL	$27.7^{\pm 0.59}$	$21.2^{\pm 0.64}$	$23.5^{\pm 0.70}$	$12.5^{\pm 0.66}$	$30.2^{\pm 0.57}$	$22.2^{\pm 0.62}$
Entropy	$12.5^{\pm 0.51}$	$20.9^{\pm 0.64}$	$15.9^{\pm 0.70}$	$22.5^{\pm 0.73}$	$13.0^{\pm 0.52}$	$20.9^{\pm 0.65}$
IC	$9.1^{\pm 0.48}$	$20.5^{\pm 0.60}$	$9.3^{\pm 0.61}$	$17.7^{\pm 0.64}$	$8.5^{\pm 0.46}$	$14.0^{\pm 0.52}$
GLSim	$20.1^{\pm 0.47}$	$15.7^{\pm 0.49}$	$22.1^{\pm 0.66}$	$23.5^{\pm 0.70}$	$16.6^{\pm 0.50}$	$18.7^{\pm 0.57}$
SVAR	$2.7^{\pm 0.28}$	$2.2^{\pm 0.28}$	$-2.4^{\pm 0.21}$	$0.9^{\pm 0.21}$	$12.6^{\pm 0.39}$	$12.4^{\pm 0.43}$

Table 11. 95% confidence interval of AUROC gain by PAS over baselines for 7B models, supplementing Tab. 2. Positive value indicates improvement over baseline.

Method	LLaVA-1.5-13B		InstructBLIP 13B		Cambrian-1-13B	
	MSCOCO	Pascal VOC	MSCOCO	Pascal VOC	MSCOCO	Pascal VOC
NLL	$27.3^{\pm 0.59}$	$21.7^{\pm 0.62}$	$24.1^{\pm 0.72}$	$19.5^{\pm 0.67}$	$10.7^{\pm 1.05}$	$4.8^{\pm 0.81}$
Entropy	$12.6^{\pm 0.55}$	$23.4^{\pm 0.66}$	$8.9^{\pm 0.67}$	$11.3^{\pm 0.65}$	$25.9^{\pm 1.04}$	$34.4^{\pm 0.94}$
IC	$10.5^{\pm 0.51}$	$19.0^{\pm 0.55}$	$12.5^{\pm 0.67}$	$21.2^{\pm 0.63}$	$12.7^{\pm 0.96}$	$18.5^{\pm 0.79}$
GLSim	$9.2^{\pm 0.42}$	$4.9^{\pm 0.35}$	$1.8^{\pm 0.43}$	$2.7^{\pm 0.40}$	$-0.2^{\pm 0.72}$	$-1.7^{\pm 0.49}$
SVAR	$1.8^{\pm 0.26}$	$2.6^{\pm 0.26}$	$1.3^{\pm 0.23}$	$2.2^{\pm 0.25}$	$3.7^{\pm 0.63}$	$5.6^{\pm 0.55}$

Table 12. 95% confidence interval of AUROC gain by PAS over baselines for 13B models, supplementing Tab. 6. Positive value indicates improvement over baseline.