NEURAL INFORMATION PROCESSING SYSTEMS

Intelligent Comput. imaging Lab.

On Epistemic Uncertainty of Visual Tokens for Object Hallucinations in Large Vision-Language Models

Hoigi Seo[1*], Dong Un Kang[1*],

Hyunjin Cho[1], Joohoon Lee[2], and Se Young Chun[1,2,3†]

* Authors contributed equally, † Corresponding author

[1]Dept. Of Electrical and Computer Engineering, [2]INMC & [3]IPAI

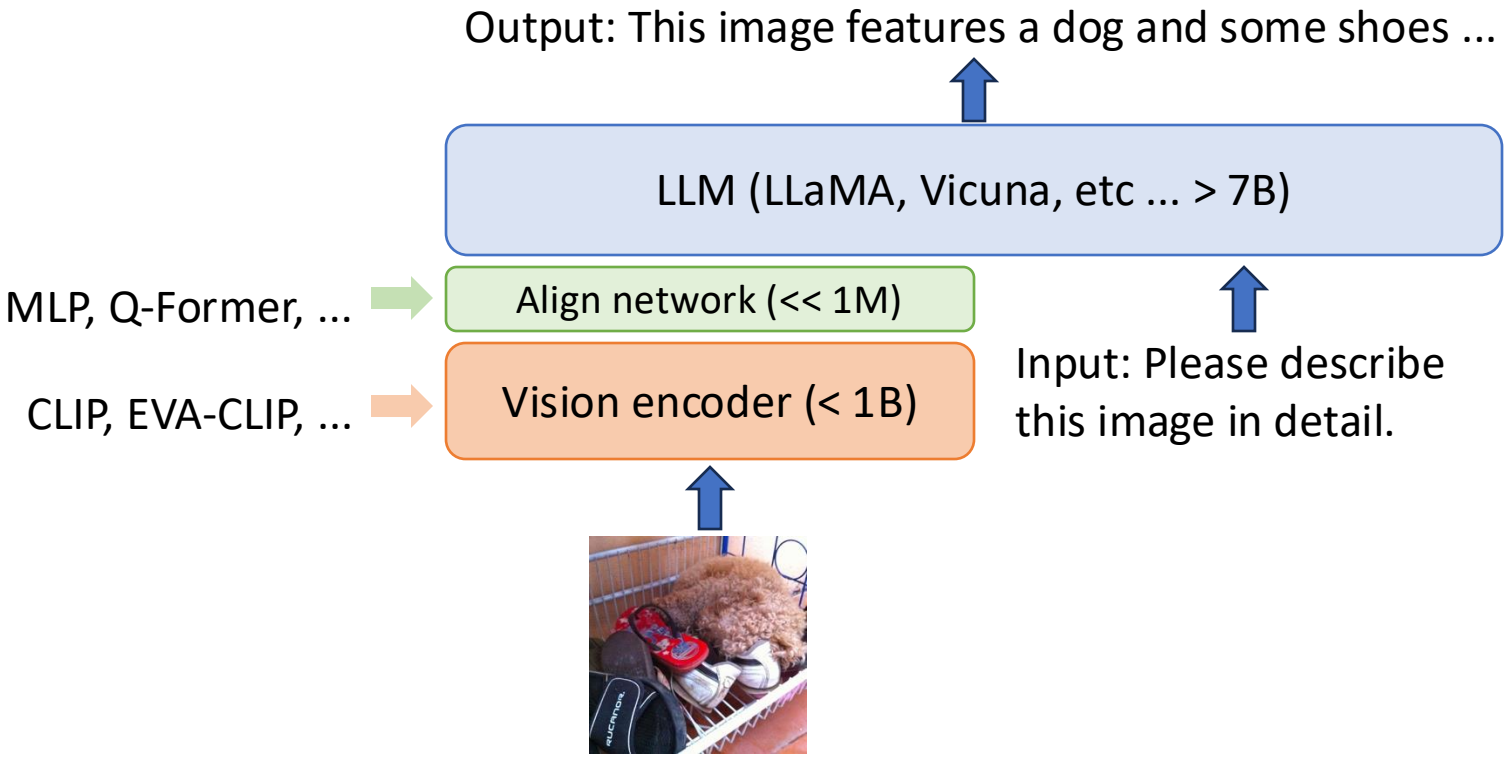
Seoul National University, Republic of Korea

{seohoiki3215, qkrtnskfk23, jim0228, joohoonl, sychun}@snu.ac.kr

Large Vision-Language Models (LVLMs)

☐ Large Vision-Language Models

- Incorporates pre-trained vision encoder with Large Language Models (LLM) to handle multi-modal inputs
- Usually consists of vision encoder, multi-modal aligner, language model.



Hallucinations in LVLMs

☐ Hallucinations

- A phenomenon in which incorrect information is output
- Object Hallucination: A phenomenon in which an object that does not exist in the input image given in LVLM is considered to exist.

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.

Prompt: Describe this photo in detail.

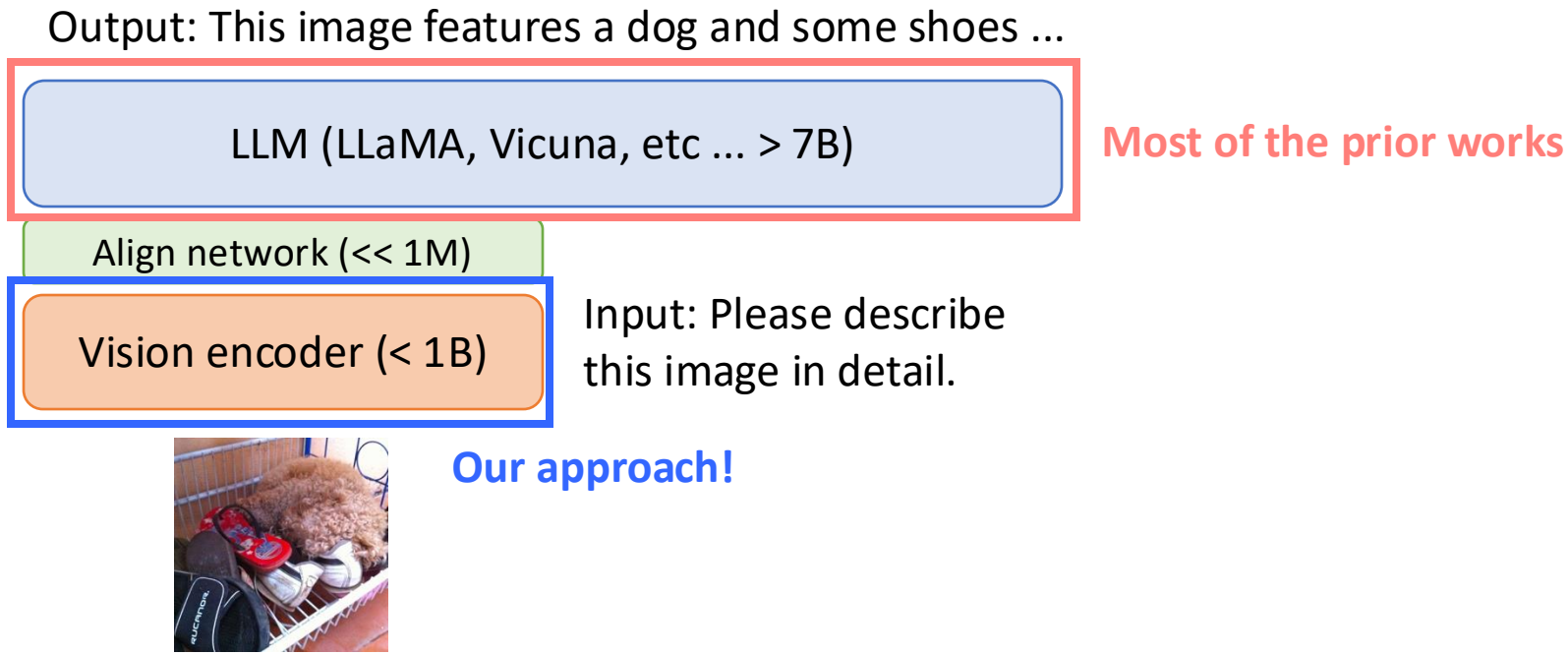
Regular Decoding: In the image, a man is standing on the back of a Subaru in a city with the color of a Ford SUV parked in the background. He appears to be ironing clothes while standing on a wine rack located under the vehicle. He is wearing a yellow shirt. There is a **white and blue taxi** parked nearby the yellow Subaru. Additionally, there are several buildings located in the cityscape, perhaps in the background of the Subaru and the taxi, and possibly overlapping with the yellow Subaru.

Leng, S., Zhang, H., Chen, G., Li, X., Lu, S., Miao, C., & Bing, L. (2024). Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (pp. 13872-13882).

Mitigating hallucinations in LVLMs

☐ Cause of the object hallucinations in LVLMs

- Multiple perspectives have been proposed to explain the causes of object hallucination — Mostly on language side.
  - Language prior / Vision-Language misalignment / decoding errors ...
- We identify a new cause — the epistemic uncertainty of the vision encoder.



Preliminary – Epistemic uncertainty

☐ Epistemic vs. Aleatoric uncertainty

- There are two types of uncertainty
- Epistemic uncertainty
  - Uncertainty caused by model parameters – measurements of how confident our model is in the results.
  - Epistemic uncertainty can be reduced with model training / post-processing
- Aleatoric uncertainty
  - Uncertainty caused by the data itself – uncertainty arising from the inherent noise in the data.
  - It is common to assume heteroscedastic uncertainty, where noise is different for each input, and it does not disappear even if you acquire a lot of data. This is an uncertainty that we cannot generally mitigate.

Preliminary – Adversarial attack

☐ Adversarial Attack on LVLM

- Inducing hallucination through small perturbations in the image
- Most of them go up to the LLM and perform noise optimization.
- Breaking the Visual Perception (Wang et al.) shows that attacks on LVLM are successful even when attacking only the image encoder.

☐ Projected Gradient Descent

- Most popular method for adversarial attack
- Optimization is performed by applying projection operation so that perturbation exists inside the epsilon ball.

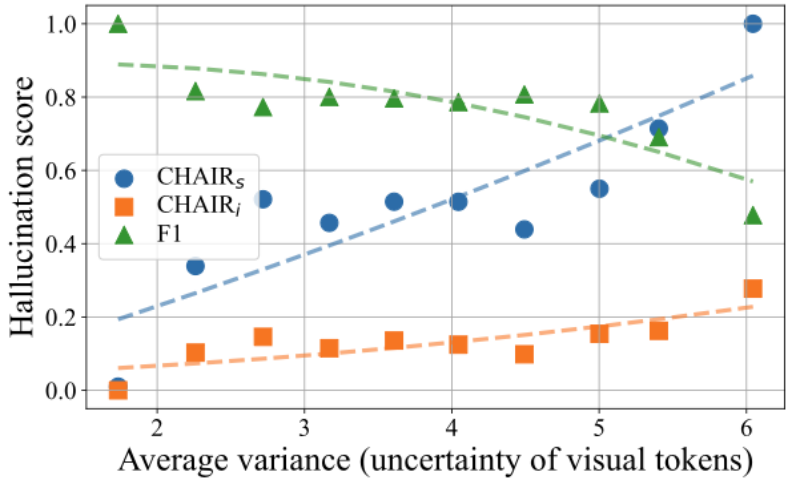
$$X_{i+1} = \Pi(X_i + \text{sign}(\nabla_{X_i}\mathcal{L}(\text{sim}(F(X_i), F(X_0)))))$$

Wang, Y., Liu, C., Qu, Y., Cao, H., Jiang, D., & Xu, L. (2024, October). Break the visual perception: Adversarial attacks targeting encoded visual tokens of large vision-language models. In *Proceedings of the 32nd ACM International Conference on Multimedia* (pp. 1072-1081).

Analysis

☐ Uncertain visual tokens contribute to object hallucination

- We measure the epistemic uncertainty of each visual token via Monte-Carlo dropout.
- We report the average variance and hallucination score of each group by binning visual tokens with uncertainty of the top 1-sigma.
- The Spearman's rank correlation test results show that $\mathrm{CHAIR_S}$, $\mathrm{CHAIR_i}$, and F1 all have p-values < 0.05, indicating a *'strong'* increasing/decreasing relationship.



Upper bound on the differential entropy

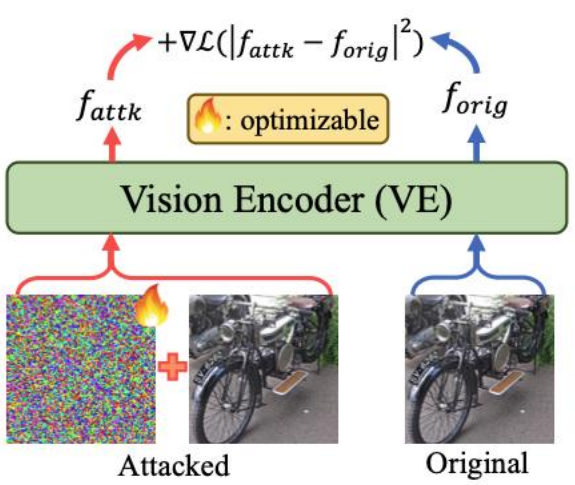
☐ Adversarial attack can efficiently approximates the uncertainty.

- The most common measure of epistemic uncertainty is Monte-Carlo dropout performed over thousands of forward passes.
  - However, this requires too much runtime and computational cost.
- We prove, through the following lemma and theorem, that an adversarial attack of hundreds of iterations implies an upper bound on the differential entropy of each visual token.

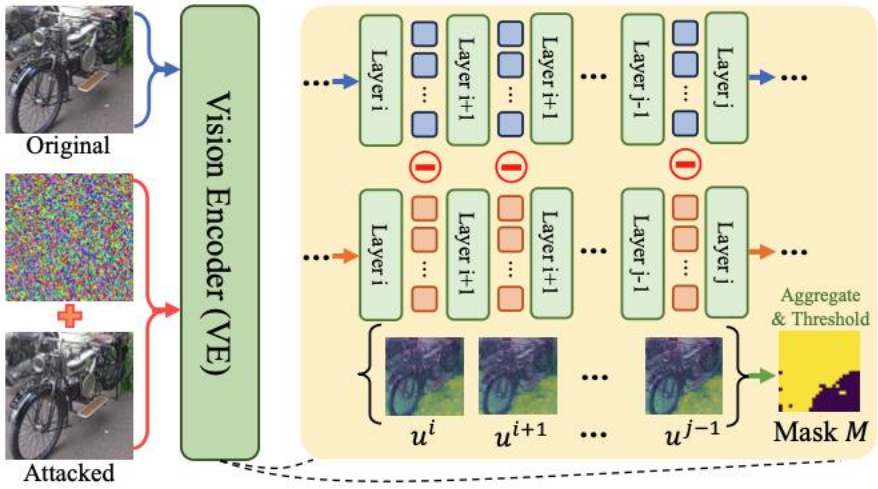
Theorem 3.2 (Upper bound of differential entropy increases as hidden state deviation increases under adversarial attack)**.** *Let x be an input image, and let ϵ be a small adversarial perturbation. Define the perturbed input as $X := x + \epsilon$. Let $f = \{f_t\}_{t=1}^{L}$ be a smooth L-block transformer that processes a sequence of N input tokens. Let $z^{(t)} := f_t \circ \cdots \circ f_1(x) \in \mathbb{R}^{N \times d}$ and $Z^{(t)} := f_t \circ \cdots \circ f_1(X) \in \mathbb{R}^{N \times d}$ be the hidden states at layer t for the clean and perturbed inputs, respectively. Denote the i-th token representation at layer t as $z_i^{(t)} \in \mathbb{R}^d$ and $Z_i^{(t)} \in \mathbb{R}^d$. If $Z_i^{(t)}$ changes smoothly with small ϵ, then the upper bound of the differential entropy of $Z_i^{(t)}$ increases as $\mathbb{E}_\epsilon[\|Z_i^{(t)} - z_i^{(t)}\|_2^2]$ increases.*

Uncertainty mask generation

☐ Adversarial attack & Uncertainty mask generation



(a) Adversarial attack process



(b) Uncertainty mask generation process

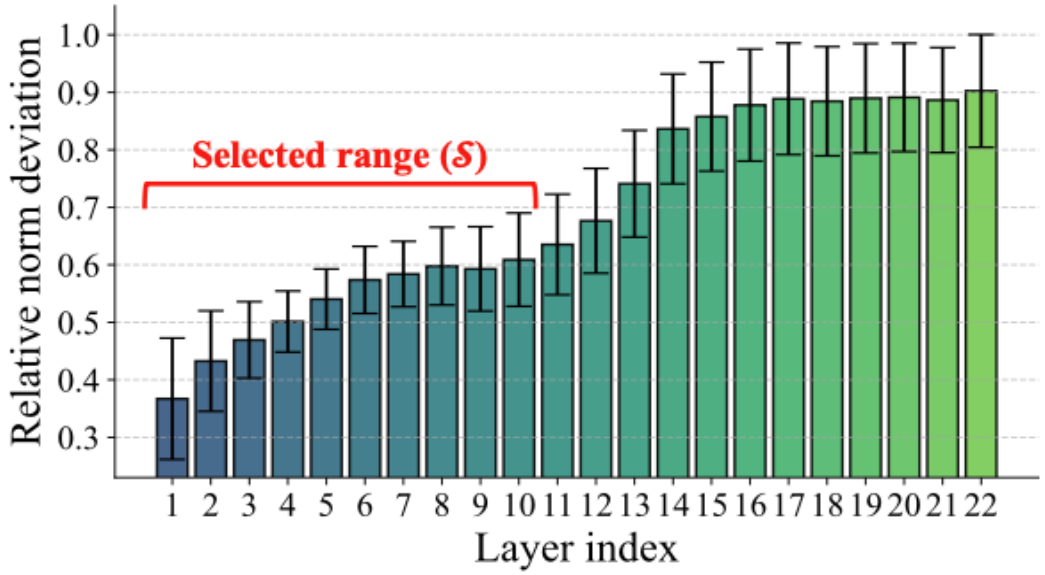
$$U = \frac{1}{j-i}\sum_{l=i}^{j-1}\frac{u^l - u^l_{\min}}{u^l_{\max} - u^l_{\min}}.$$

$$M = 1 - \frac{1}{2}\left\lfloor \operatorname{sign}\left(\left(\frac{U-\mu_U}{\sigma_U}\right) - \sigma_{\text{th}}\right) + 1 \right\rfloor \in \{0,1\}^N.$$

Verification on layer selection

☐ Selecting the uncertainty map extraction layer

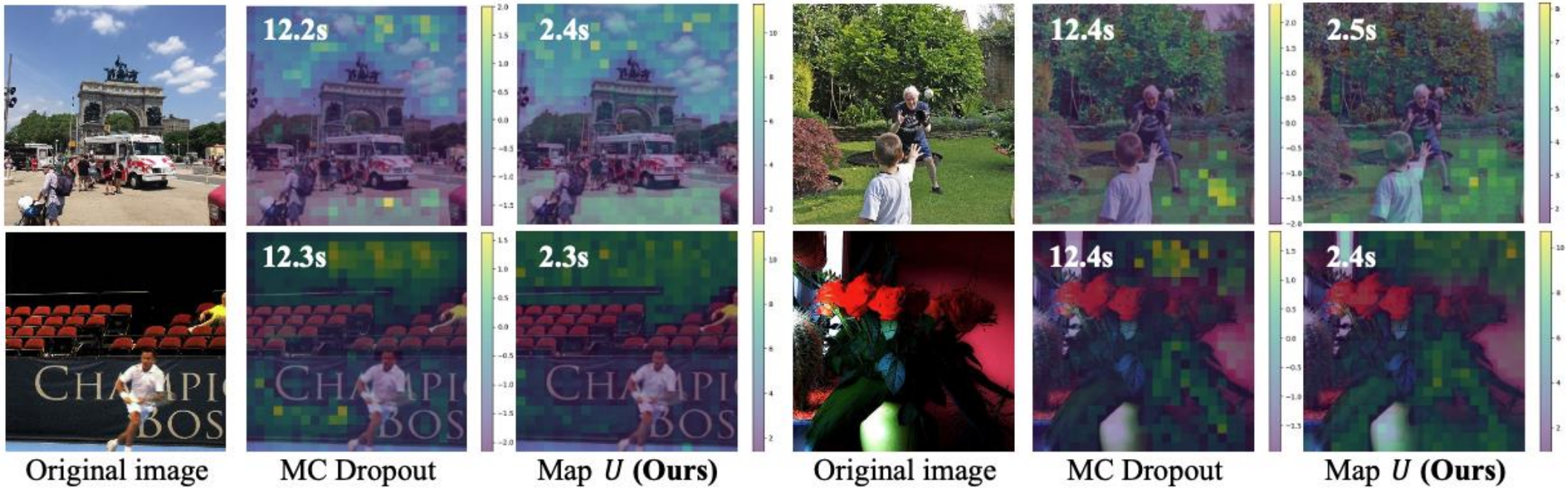
- We made an assumption in the theorem that the change in hidden state is small.
- When we checked the amount of change in hidden states in each layer, the change in the early layer was smaller.
- Therefore, we extracted maps from early layers, which is consistent with the results from the ablation study.



Verification on similarity with MC dropout

☐ Does the uncertainty map we obtained really match the uncertainty map obtained by MC dropout?

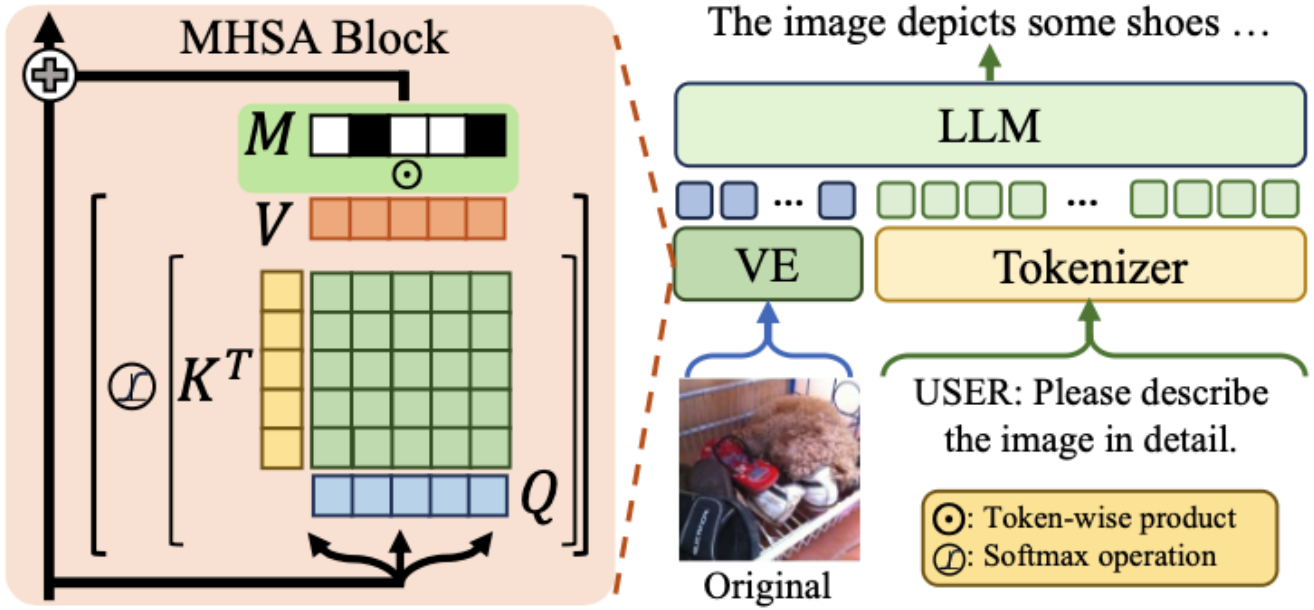
- Yes!
- MC dropout: 1,000 forward passes / Ours: 100 PGD attack iteration



Masking uncertain visual tokens

☐ Masking uncertain visual tokens for training-free hallucination mitigation

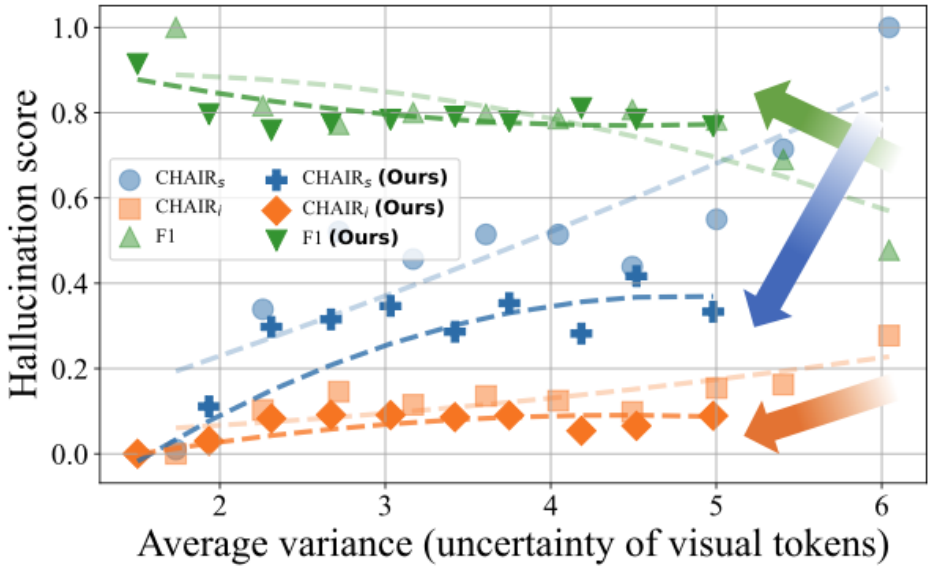
- With the uncertainty mask M we've obtained in previous stage, we mask the *"uncertain"* visual tokens in the middle-layer of self-attention operation



Verification on uncertainty reduction

☐ Does our method reduce uncertainty and mitigate object hallucination?

- Yes!
- We performed a Wilcoxon signed rank test and confirmed statistical significance with a p-value < 0.05, indicating that uncertainty decreased and $\mathrm{CHAIR_S}$ and $\mathrm{CHAIR_I}$ decreased.



Benchmarks – CHAIR

☐ CHAIR

- After asking LVLM to generate a description for the input image, any object that does not exist in the image is considered a hallucination.
  - 500 images + caption pairs
- $\mathrm{CHAIR_S}$ and $\mathrm{CHAIR_i}$ measure hallucination at sentence and image levels, respectively. (The lower the better)
- F1 score measures the quality of the generated description. (The higher the better)

Image Model predictions:
bowl, broccoli, carrot, dining table

Language Model predictions for the last word:
fork, spoon, bowl

Generated caption: A plate of food with broccoli and a ***fork***.

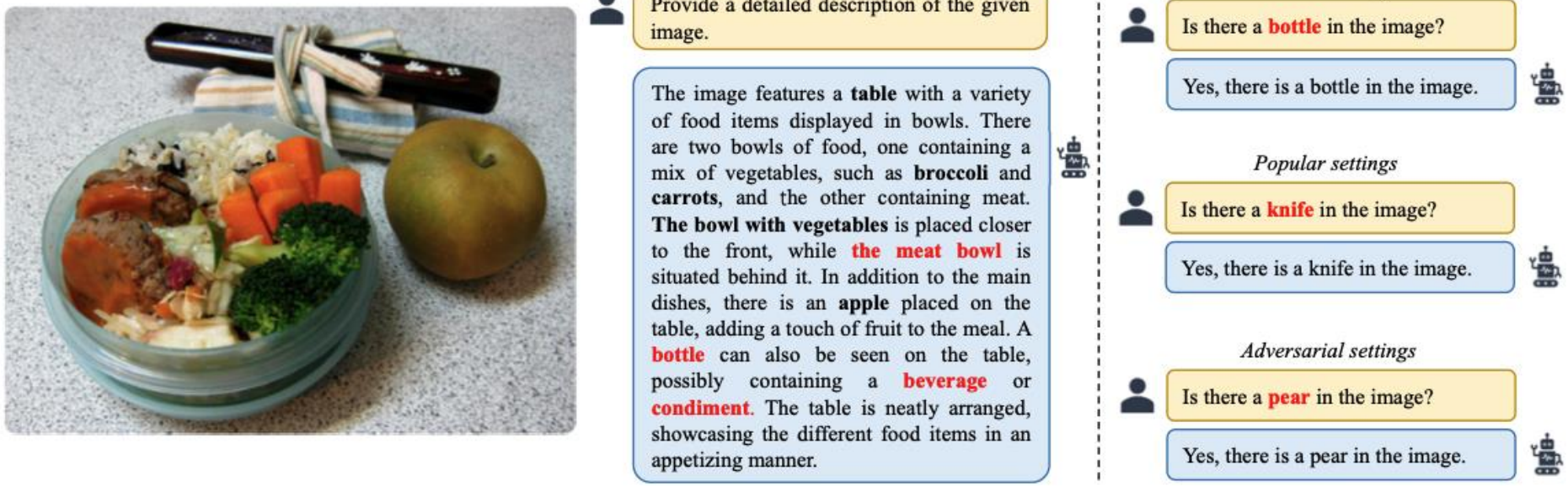
Figure 2: Example of image and language consistency. The hallucination error ("fork") is more consistent with the Language Model.

Rohrbach, A., Hendricks, L. A., Burns, K., Darrell, T., & Saenko, K. (2018). Object hallucination in image captioning. arXiv preprint arXiv:1809.02156.

Benchmarks – POPE

☐ POPE

- Ask LVLM whether an object exists in a given image with a Yes/No question and measure accuracy.
- It consists of 3 sets: Random / Popular / Adversarial, and each has 3,000 prompts → total 9,000 prompts

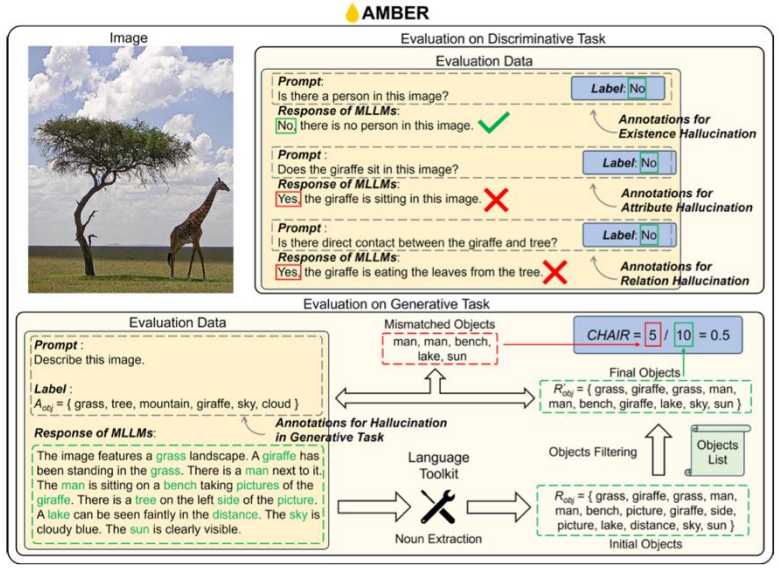


Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, X., & Wen, J. R. Evaluating Object Hallucination in Large Vision-Language Models. In The 2023 Conference on Empirical Methods in Natural Language Processing.

Benchmarks - AMBER

☐ AMBER

- It is divided into a generative task similar to CHAIR and a discriminative task similar to POPE.
- To measure object hallucination, the discriminative task uses the ‘existence’ subset, and the generative task uses the full set.
- Consists of 5,928 prompts



Wang, J., Wang, Y., Xu, G., Zhang, J., Gu, Y., Jia, H., ... & Sang, J. (2023). Amber: An llm-free multi-dimensional benchmark for mllms hallucination evaluation. arXiv preprint arXiv:2311.07397.

Quantitative results – CHAIR and POPE

	Method	Greedy			OPERA			VCD			PAI			Devils		
		Orig.	+Ours	Δ%	Orig.	+Ours	Δ%	Orig.	+Ours	Δ%	Orig.	+Ours	Δ%	Orig.	+Ours	Δ%
LLaVA-1.5-7B	$C_s \downarrow$	47.4	29.2	↓38.4%	47.8	29.4	↓38.5%	53.8	35.2	↓34.6%	33.2	26.0	↓21.7%	27.0	**23.0**	↓14.8%
	$C_i \downarrow$	12.2	9.3	↓23.8%	12.8	9.5	↓25.8%	15.2	10.7	↓29.6%	8.5	7.9	↓7.1%	6.6	**5.6**	↓15.2%
	F1 ↑	77.9	78.2	↑0.4%	77.7	**78.4**	↑0.9%	75.2	75.2	↑0.0%	78.3	77.2	↓1.4%	78.3	78.0	↓0.4%
	Rand. ↑	89.3	89.3	↑0.0%	89.2	88.6	↓0.7%	84.6	86.2	↑1.9%	89.4	89.2	↓0.2%	89.6	**90.0**	↑0.4%
	Pop. ↑	85.8	85.8	↑0.0%	85.8	85.2	↓0.7%	82.4	82.9	↑0.6%	86.0	86.4	↑0.5%	86.4	**87.2**	↑0.9%
	Adv. ↑	79.3	80.0	↑0.9%	**80.3**	79.6	↓0.9%	77.0	78.1	↑1.4%	79.5	79.9	↑0.5%	78.6	79.6	↑1.3%
Shikra-7B	$C_s \downarrow$	58.0	43.2	↓25.5%	34.8	28.8	↓17.2%	56.2	47.2	↓16.0%	32.4	22.2	↓31.5%	24.4	**20.6**	↓15.6%
	$C_i \downarrow$	15.6	11.7	↓25.0%	11.1	9.6	↓13.5%	16.1	12.8	↓20.5%	7.8	**6.1**	↓21.8%	7.6	6.8	↓10.5%
	F1 ↑	74.7	**76.9**	↑2.9%	74.2	74.2	↑0.0%	74.4	75.2	↑1.1%	76.7	75.1	↓2.1%	73.3	72.2	↓1.5%
	Rand. ↑	83.2	85.1	↑2.3%	84.8	**85.4**	↑0.7%	82.1	82.7	↑0.7%	83.9	84.0	↑0.1%	83.8	82.5	↓1.6%
	Pop. ↑	82.3	82.6	↑0.4%	82.8	82.1	↓0.8%	79.7	80.7	↑1.3%	**83.1**	80.7	↓2.9%	79.9	78.2	↓2.1%
	Adv. ↑	78.2	78.8	↑0.8%	79.2	**79.7**	↑0.6%	77.3	77.1	↓0.3%	78.8	77.4	↓1.8%	77.7	76.7	↓1.3%
MiniGPT-4	$C_s \downarrow$	28.6	27.4	↓4.2%	23.8	22.6	↓5.0%	32.0	30.6	↓4.4%	19.6	**17.8**	↓9.2%	21.6	20.8	↓3.7%
	$C_i \downarrow$	8.5	8.3	↓2.4%	8.8	8.5	↓3.4%	9.7	9.1	↓6.2%	6.2	**6.0**	↓3.2%	7.5	7.0	↓6.7%
	F1 ↑	71.5	71.3	↓0.3%	69.8	70.0	↑0.3%	70.2	71.3	↑1.7%	**71.7**	**71.7**	↑0.0%	70.1	70.4	↑0.4%
	Rand. ↑	**82.8**	82.5	↓0.4%	74.2	74.4	↑0.3%	59.2	59.3	↑0.2%	82.1	82.0	↓0.1%	77.4	77.8	↑0.5%
	Pop. ↑	75.1	74.6	↓0.7%	71.3	71.8	↑0.7%	54.9	55.0	↑0.2%	**75.8**	75.2	↓0.8%	68.4	68.6	↑0.3%
	Adv. ↑	71.8	71.2	↓0.8%	69.7	69.4	↓0.4%	53.8	54.2	↑1.1%	**72.1**	71.6	↓0.7%	65.2	65.3	↑0.2%

Quantitative results - AMBER

	Method	Greedy			OPERA			VCD			PAI			Devils		
		Orig.	+Ours	Δ%	Orig.	+Ours	Δ%	Orig.	+Ours	Δ%	Orig.	+Ours	Δ%	Orig.	+Ours	Δ%
Gen.	CHAIR ↓	6.7	5.1	↓23.9%	7.4	5.8	↓21.6%	8.5	6.1	↓28.2%	5.1	4.7	↓7.8%	4.1	**3.9**	↓4.9%
	Hal ↓	30.2	24.2	↓19.9%	33.0	23.3	↓29.4%	38.4	28.6	↓25.5%	25.1	22.5	↓10.4%	21.0	**20.9**	↓0.5%
	Cog ↓	3.8	2.3	↓39.5%	3.7	2.1	↓43.2%	4.4	2.3	↓47.7%	1.9	1.9	↓0.0%	**1.4**	1.5	↑7.1%
Disc.	Pre. ↑	100.0	100.0	↑0.0%	100.0	100.0	↑0.0%	100.0	100.0	↑0.0%	100.0	100.0	↑0.0%	100.0	100.0	↑0.0%
	Rec. ↑	71.2	78.0	↑9.6%	74.9	**81.0**	↑7.5%	67.3	75.7	↑12.5%	71.9	74.1	↑3.1%	72.5	75.2	↑3.7%
	F1 ↑	83.2	87.6	↑5.3%	85.6	**89.5**	↑4.6%	80.4	86.2	↑7.2%	83.6	85.1	↑1.8%	84.1	85.8	↑2.0%
	AMBER↑	88.2	91.2	↑3.4%	89.1	**91.8**	↑3.0%	86.0	90.1	↑4.8%	89.2	90.2	↑1.1%	90.0	91.0	↑1.1%

Qualitative results

USER: Please describe the image in detail.

LLaVA: The image features a train station with two trains on the tracks. The red train is positioned closer to the left side of the image, while the white and blue train is situated further back. There are **several people** in the scene, with some standing near the trains and others scattered throughout the station. A **bench** can be seen in the middle of the station, providing a place for passengers to sit. Additionally, there is a **handbag** placed on the ground near the center of the scene, possibly belonging to one of the **passengers**.

LLaVA + **Ours**: The image features a red train traveling down the tracks, with another train visible in the background. The red train is positioned in the middle of the scene, while the other train is located towards the right side of the image. The trains are both moving along the tracks, creating a sense of motion and activity.

Input image

USER: Please describe the image in detail.

Shikra: The image features a one-way street sign prominently placed next to a white brick building. A bench is situated close to the street sign, providing a place for pedestrians to sit and observe the surroundings. The scene also includes **a few cars** parked along the street, with one **car** closer to the left side of the image, another **car** further back, and a third **car** on the right side of the scene. Additionally, there is **a fire hydrant** visible on the right side of the image, near the edge of the frame.

Shikra + **Ours**: The image features a one-way street sign positioned in front of a building. The sign is pointing to the left, indicating that vehicles should travel in one direction on this particular street. Beside the sign, a bench can be seen, providing a place for pedestrians to sit and observe the surroundings. The overall scene suggests a view of a street corner with some graffiti on the wall behind the sign.

Input image

Ablation study

Table 3: **Impact of vision encoder layers on generating the uncertainty mask M.** Using early layers of vision encoder (1–10) to compute M yields the most effective object hallucination mitigation performance.

Mask Source Layer	$C_s \downarrow$	$C_i \downarrow$	F1 ↑
Greedy	47.4	12.2	77.9
Layers 1–10	**29.2**	**9.3**	**78.2**
Layers 11–20	44.2	12.7	77.4
Layers 21–22	41.8	12.1	77.9

Table 4: **Effect of applying the uncertainty mask M to different layers in the vision encoder.** Applying the mask at middle layers of vision encoder (13–17) results in the most effective performance.

Masking Layer Range	$C_s \downarrow$	$C_i \downarrow$	F1 ↑
Greedy	47.4	12.2	77.9
Layers 1–8	45.0	12.6	77.9
Layers 8–12	55.8	15.5	75.7
Layers 13–17	**29.2**	**9.3**	**78.2**
Layers 18–22	45.8	13.0	77.7

Table 5: **Comparison of masking strategies for uncertain visual tokens.** We compare our attention-level masking method with alternatives applied at different stages of the vision encoder (VE). S.M. denotes soft masking, which attenuates uncertain tokens by a small factor (e.g., 0.1 or 0.2).

Strategy	Greedy	Input of VE	Output of VE	MLP Layer	S.M. (0.1 / 0.2)	**Ours**
$C_s \downarrow$	47.4	47.4	34.4	51.0	35.0 / 40.0	**29.2**
$C_i \downarrow$	12.2	12.5	10.0	13.5	10.4 / 11.5	**9.3**
F1 ↑	77.9	77.5	74.7	77.9	**78.3** / 78.1	78.2

NEURAL INFORMATION PROCESSING SYSTEMS

Thank You!

Intelligent Comput. imaging Lab.

On Epistemic Uncertainty of Visual Tokens for Object Hallucinations in Large Vision-Language Models

For more results and code, please visit our project page!

https://keenyjin.github.io/epistemic





Hoigi Seo[1*] **Dong Un Kang[1*]** **Hyunjin Cho[1]** **Joohoon Lee[2]** **Se Young Chun[1,2,3†]**

* Authors contributed equally, † Corresponding author

[1]Dept. Of Electrical and Computer Engineering, [2]INMC & [3]IPAI
Seoul National University, Republic of Korea

Acknowledgements

This work was supported in part by Institute of Information \& communications Technology Planning \& Evaluation (IITP) grants funded by the Korea government(MSIT) [NO.RS-2021-II211343, Artificial Intelligence Graduate School Program (Seoul National University)], (No.RS-2025-02314125, Effective Human-Machine Teaming With Multimodal Hazy Oracle Models) and by the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT) (No. RS-2025-02263628). Also, the authors acknowledged the financial support from the BK21 FOUR program of the Education and Research Program for Future ICT Pioneers, Seoul National University.