



南京航空航天大学

Nanjing University of Aeronautics and Astronautics



模式分析与机器智能
工业和信息化部重点实验室

MIT Key Laboratory of
Pattern Analysis & Machine Intelligence

Where MLLMs Attend and What They Rely On: Explaining Autoregressive Token Generation

**Ruoyu Chen^{1,2}, Xiaoqing Guo³, Kangwei Liu^{1,2}, Siyuan Liang⁴, Shiming Liu⁵,
Qunli Zhang⁵, Laiyuan Wang⁶, Hua Zhang^{1,2,✉}, Xiaochun Cao^{7,✉}**

¹Institute of Information Engineering, Chinese Academy of Sciences ²University of Chinese Academy of Sciences

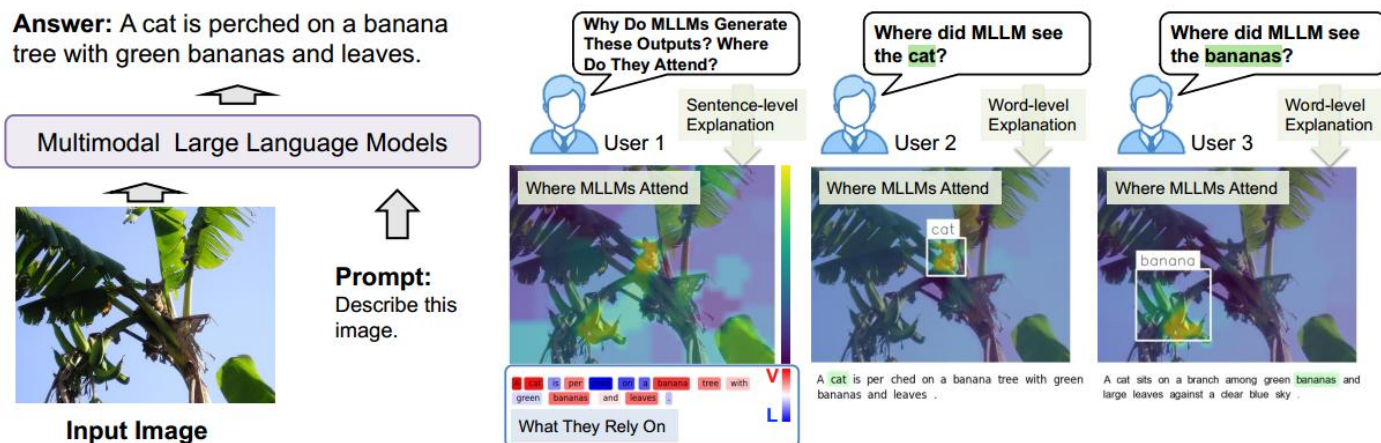
³Department of Computer Science, Hong Kong Baptist University ⁴College of Computing and Data Science, NTU ⁵Huawei

⁶School of Flexible Electronics, SYSU ⁷School of Cyber Science and Technology, Shenzhen Campus of Sun Yat-sen University

CVPR 2026

◦ Main issue

1. When MLLM generates an answer, which areas of the image actually drive the appearance of a certain token?
2. Does the same token rely more on visual evidence or on language priors?
3. If the model hallucinates, is it possible to identify the smallest visual region that triggers the hallucination?



◦ Comparisons

- ▶ “Activate” or “Attention” more like “Preferences” — Not faithful
- ▶ Gradient method is intensive and is susceptible to noise in long sequences and modal imbalance.
- ▶ Difficult to jointly support: sentence-level, word-level explanations, hallucination diagnosis.

Motivation

Outline

Divide the image into superpixel subregions, **then** identify the set of regions that is “as small as possible while still best supporting the generation of the target token”; **then** check whether the target token drops significantly after removing these regions.

Insight 1

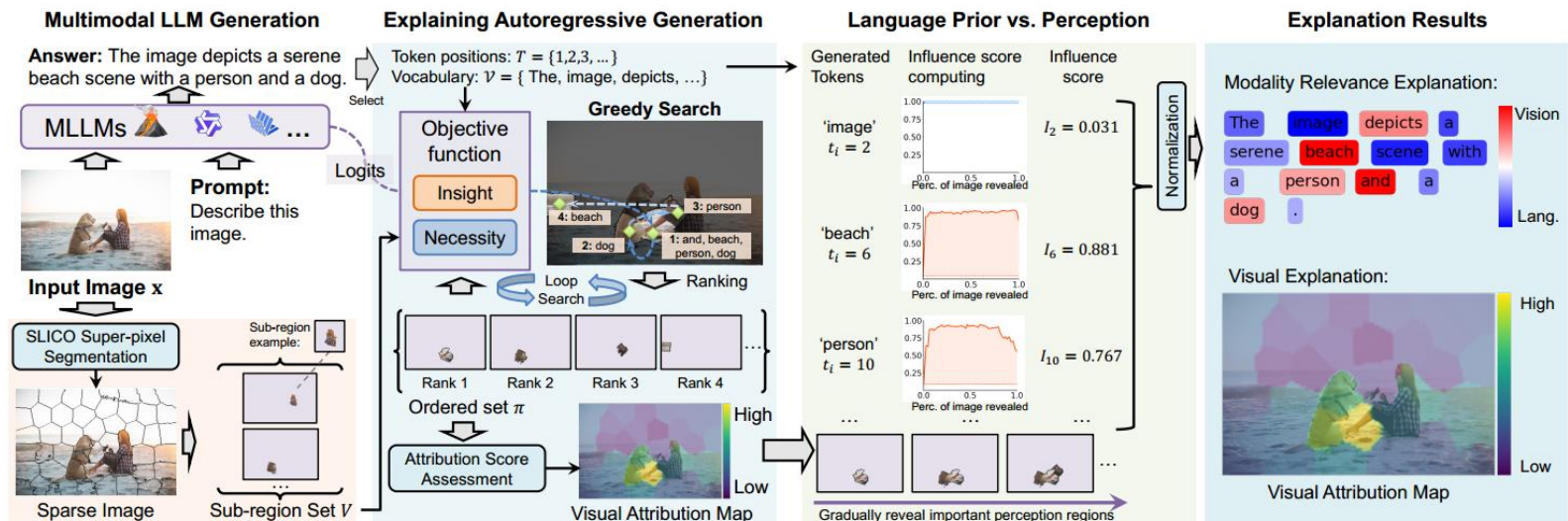
A lightweight black-box framework that does not rely on internal model gradients or activation functions.

Insight 2

A combined objective that jointly constrains both “sufficiency” and “necessity.”

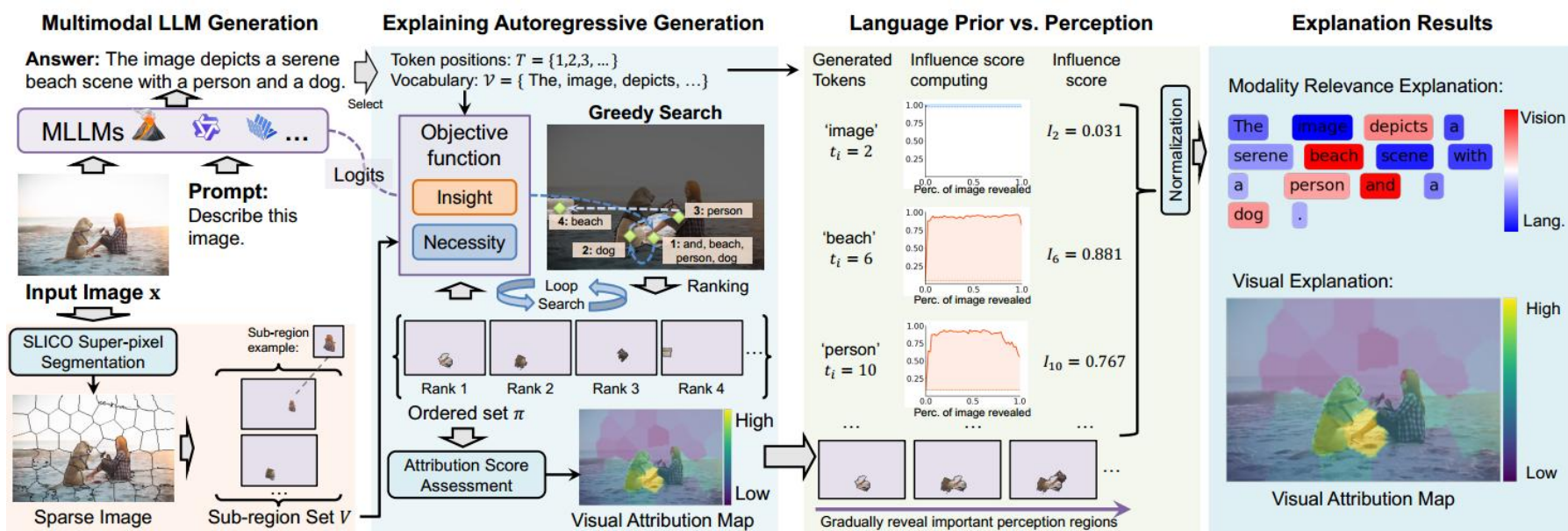
Insight 3

Influence score, distinguishing the relative contributions of language priors and visual evidence.



From Image Subregions to Ordered Attribution

- **Step 1:** Split the input image into a set of subregions V using **SLICO**
- **Step 2:** Reframe the attribution problem as subset selection / ordered subset optimization
- **Step 3:** Use greedy search to identify the most critical regions and rank them by importance π
- **Step 4:** Based on the sorting, further analyze the strength of visual dependencies for each token



SLICO → Objective → Greedy Search → Influence Score

$$\max_{\pi \in \mathcal{P}(V), |\pi| < k} \sum_{r=1}^{|\pi|} \mathcal{F}(\pi, r)$$

Objective: Insight and Necessity

Insight asks, “Is having it enough?”; Necessity asks, “Can we manage without it?”

Insight score

If only the candidate regions S are retained, is the probability of the target token still high?
If so, it indicates that this set of regions is sufficient to support the generation.

Necessity score

Does the probability of the target token decrease significantly after removing region S ?
If so, this indicates that these regions are essential for generation.

Greedy ranking

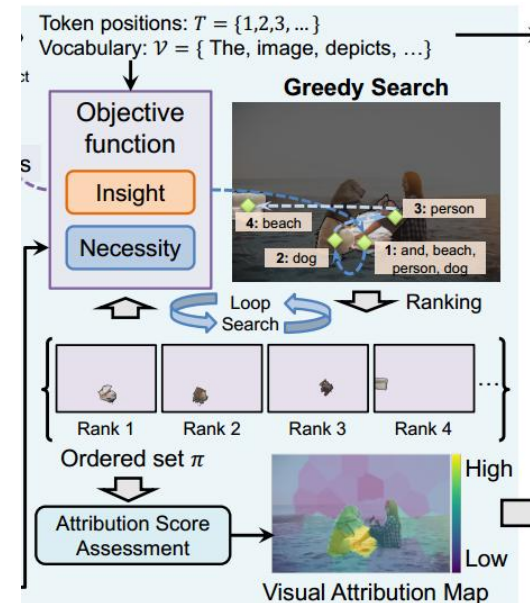
Each step identifies the area that yields the greatest marginal gain, so in the end, there are not only “critical areas” but also a “critical sequence.”

$$\mathcal{F}(S) = s_{\text{insight}}(S) + s_{\text{necessity}}(V \setminus S)$$

$$\max_{\pi \in \mathcal{P}(V), |\pi| < k} \sum_{r=1}^{|\pi|} \mathcal{F}(\pi_{:r})$$

$$s_{\text{insight}}(S) = \sum_i p(y_{t_i} = v_i \mid S, \text{Prompt}, y_{<t_i})$$

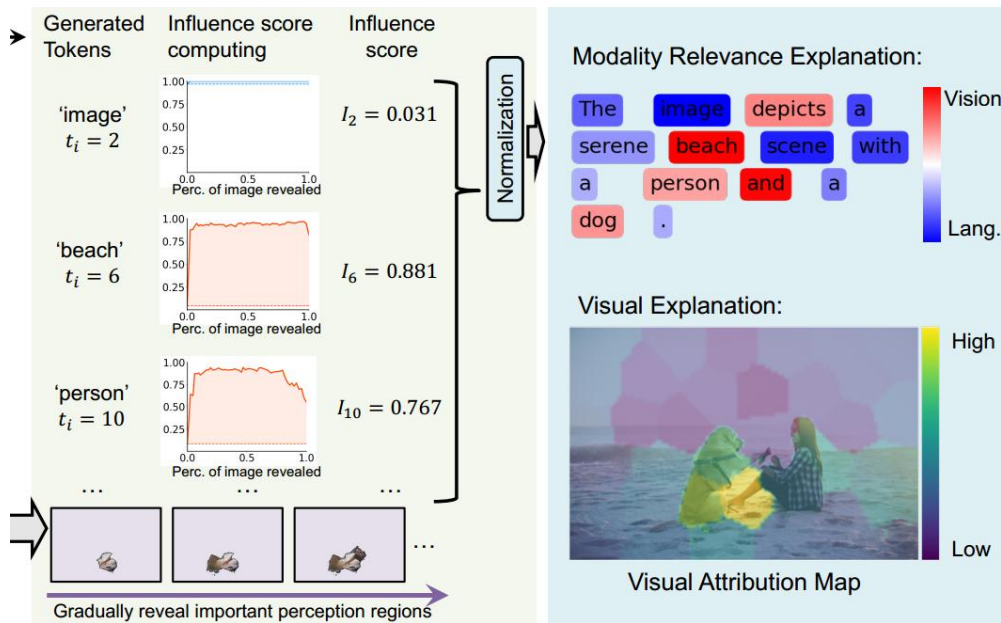
$$s_{\text{necessity}}(V \setminus S) = \sum_i \left(1 - p(y_{t_i} = v_i \mid V \setminus S, \text{Prompt}, y_{<t_i})\right)$$



What They Rely On?

EAGLE uses the ordered sequence of regions π that has already been obtained to gradually reveal important visual regions and observe how the token probability curves change.

$$I_{t_i} = \sum_{r=1}^{|\pi|} \left(p(y_{t_i} = v_i \mid \pi_{:r}, \text{Prompt}, y_{<t_i}) - \min_j p(y_{t_i} = v_i \mid \pi_{:j}, \text{Prompt}, y_{<t_i}) \right)$$



Influence score

Higher scores: the token relies more on visual evidence;
Lower scores: the token relies more on language prior knowledge.

Why is it vital?

Beyond merely asking “where the model is looking,” further explore whether this conclusion is derived from the image itself.

Experiment

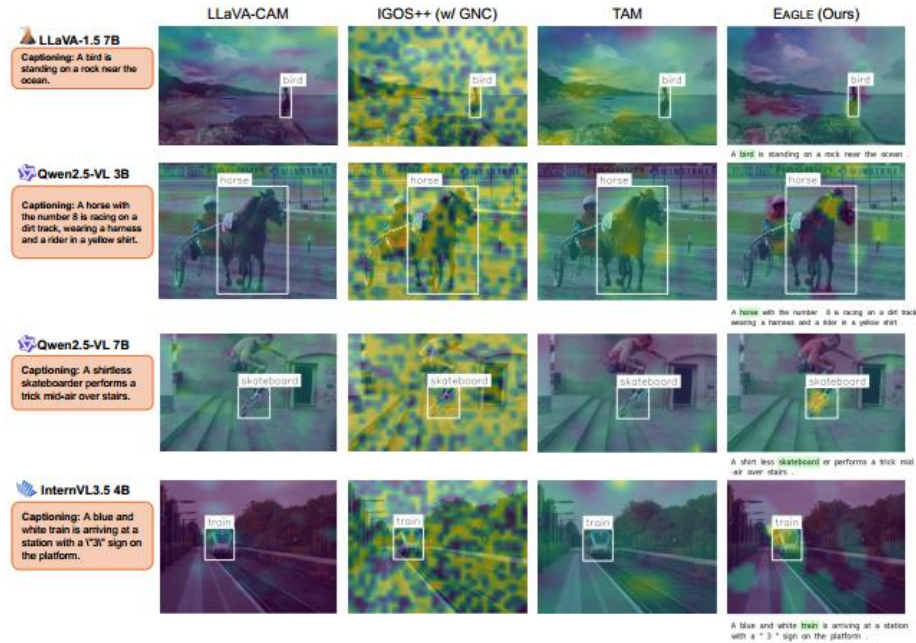


◦ Sentence-level

Datasets	MLLMs	Methods	Sentence-level Faithfulness			Sensitive Tokens-level Faithfulness			GPU Memory (↓)	Time Consum. (↓)
			Ins. (↑)	Del. (↓)	Ave. high. score (↑)	Ins. (↑)	Del. (↓)	Ave. high. score (↑)		
MS COCO [22] (Image caption task)	LLaVA-1.5 7B [24]	LLaVA-CAM [38]	0.5298	0.5317	0.6031	0.4124	0.4115	0.5783	37.25 GB	15.4 s
		iGOS++ (w/ GNC) [34]	0.5293	0.5168	0.6004	0.4101	0.3815	0.5731	48.18 GB	36.3 s
		EAGLE	0.5970	0.4554	0.6259	0.5344	0.2809	0.5993	16.07 GB	258.4 s
	Qwen2.5-VL 3B [4]	LLaVA-CAM [38]	0.4978	0.5562	0.6662	0.3541	0.4497	0.6424	28.99 GB	22.2 s
		iGOS++ (w/ GNC) [34]	0.5328	0.4891	0.6672	0.4021	0.3273	0.6473	71.62 GB	46.6 s
		EAGLE	0.6479	0.4345	0.7039	0.5867	0.2710	0.6840	8.75 GB	418.3 s
	Qwen2.5-VL 7B [4]	LLaVA-CAM [38]	0.5605	0.5464	0.7235	0.4467	0.4209	0.7010	47.17 GB	21.8 s
		iGOS++ (w/ GNC) [34]	0.5603	0.5072	0.7237	0.4400	0.3623	0.6695	96.90 GB	62.7 s
		EAGLE	0.7006	0.4597	0.7578	0.6337	0.2988	0.7285	17.68 GB	436.2 s
	InternVL3.5 4B [33]	LLaVA-CAM [38]	0.6116	0.6235	0.8032	0.4948	0.5100	0.7764	81.84 GB	54.3 s
		iGOS++ (w/ GNC) [34]	0.6271	0.5726	0.7999	0.5088	0.4337	0.7715	60.93 GB	40.1 s
		EAGLE	0.7665	0.4650	0.8335	0.7042	0.3042	0.8051	12.45 GB	803.5 s
MMVP [31] (VQA task)	LLaVA-1.5 7B [24]	LLaVA-CAM [38]	0.7756	0.7745	0.7980	0.6076	0.6044	0.7275	34.38 GB	12.2 s
		iGOS++ (w/ GNC) [34]	0.7717	0.7698	0.7965	0.5825	0.5781	0.7236	92.90 GB	34.8 s
		EAGLE	0.7960	0.7474	0.8086	0.6867	0.5027	0.7507	15.40 GB	252.6 s
	Qwen2.5-VL 3B [4]	LLaVA-CAM [38]	0.7742	0.7770	0.8181	0.5925	0.6006	0.7476	19.17 GB	16.1 s
		iGOS++ (w/ GNC) [34]	0.7719	0.7613	0.8183	0.5719	0.5356	0.7437	19.79 GB	31.8 s
		EAGLE	0.8052	0.7338	0.8339	0.6634	0.4935	0.7689	8.76 GB	222.7 s
	Qwen2.5-VL 7B [4]	LLaVA-CAM [38]	0.7505	0.7486	0.8042	0.4974	0.4847	0.7242	37.54 GB	14.5 s
		iGOS++ (w/ GNC) [34]	0.7394	0.7211	0.8036	0.4505	0.3853	0.7185	32.76 GB	31.0 s
		EAGLE	0.7824	0.6996	0.8119	0.5901	0.3675	0.7362	17.40 GB	220.8 s
	InternVL3.5 4B [33]	LLaVA-CAM [38]	0.7348	0.7458	0.8325	0.4897	0.5213	0.7575	27.20 GB	20.7 s
		iGOS++ (w/ GNC) [34]	0.7277	0.7160	0.8302	0.4743	0.4454	0.7535	62.31 GB	34.4 s
		EAGLE	0.8012	0.6782	0.8471	0.6379	0.4027	0.7762	12.26 GB	176.5 s

Experiment

- Word-level



Datasets	MLLMs	Methods	Word-level Faithfulness Metrics			Localization Metrics		GPU Memory (↓)	Time Consum. (↓)
			Insertion (↑)	Deletion (↓)	Ave. high. score (↑)	Point Game _{bbox} (↑)	Point Game _{mask} (↑)		
MS COCO [22] (Image caption task)	LLaVA-1.5 7B [24]	LLaVA-CAM [38]	0.4063	0.4035	0.6053	0.2468	0.1168	36.73 GB	17.1 s
		iGOS++ (w/ GNC) [34]	0.4093	0.3812	0.6084	0.6623	0.5584	93.12 GB	33.5 s
		TAM [17]	0.3860	0.4162	0.5988	0.1818	0.1428	16.60 GB	1.7 s
		EAGLE	0.6395	0.2047	0.7213	0.8052	0.7792	16.31 GB	283.2 s
	Qwen2.5-VL 3B [4]	LLaVA-CAM [38]	0.3417	0.4575	0.7263	0.1045	0.0621	26.01 GB	12.8 s
		iGOS++ (w/ GNC) [34]	0.4141	0.2901	0.7250	0.5822	0.4967	58.1 GB	49.9 s
		TAM [17]	0.5130	0.2797	0.7985	0.5294	0.4379	9.56 GB	4.2 s
		EAGLE	0.7353	0.1628	0.8641	0.8104	0.7745	9.22 GB	371.4 s
	Qwen2.5-VL 7B [4]	LLaVA-CAM [38]	0.4170	0.4771	0.8041	0.2176	0.1428	44.26 GB	24.5 s
		iGOS++ (w/ GNC) [34]	0.4816	0.3478	0.8080	0.6734	0.5959	82.14 GB	53.1 s
		TAM [17]	0.5768	0.3167	0.8240	0.5369	0.4060	18.75 GB	4.9 s
		EAGLE	0.8109	0.2127	0.9194	0.7785	0.7383	18.03 GB	435.6 s
InternVL3.5 4B [33]	LLaVA-CAM [38]	0.4988	0.5040	0.8588	0.3201	0.2212	81.84 GB	41.3 s	
	iGOS++ (w/ GNC) [34]	0.5192	0.3983	0.8604	0.5775	0.5181	60.06 GB	32.7 s	
	TAM [17]	0.6317	0.3517	0.8712	0.5775	0.4653	14.23 GB	4.6 s	
	EAGLE	0.8623	0.1706	0.9585	0.8052	0.7755	7.61 GB	377.3 s	

Hallucination

- Focus on the first token in the answer (Yes / No)

LLaVA-1.5 7B

Qwen2.5-VL 7B

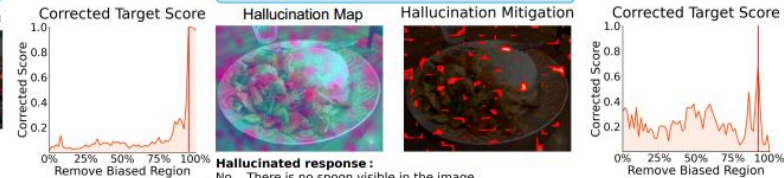
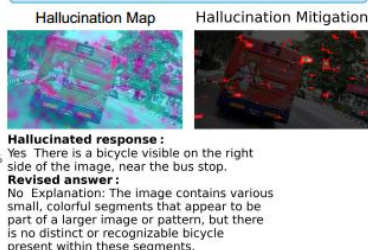
InternVL3.5 4B

IGOS++ (w/ GNC)

Question: Is there a snowboard in the image?

Question: Is there a bicycle in the image?

Question: Is there a spoon in the image?



TAM



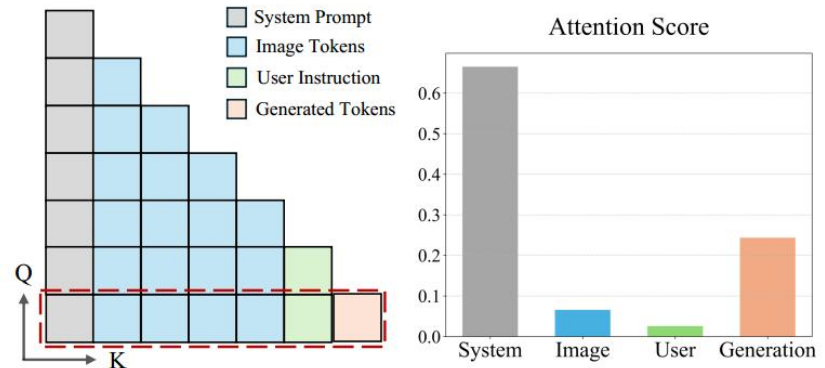
EAGLE (Ours)



- 1) Greedy search results in $O(|V|^2)$ time complexity, which explains the slow speed.
- 2) It relies on the superpixel granularity, making the partitioning process more time-consuming.
- 3) This is primarily attributed to the input space and does not yet explain the internal mechanism as a black-box.



Snowball Illusion



- Please describe this photo in detail.

- LLAVA1.5-7B:

The image features a man standing on the back of a yellow taxi cab, holding a blue shirt. He appears to be ironing the shirt while the taxi is in motion. The taxi is driving down a busy street, with other vehicles such as a **truck** and a car visible in the scene.

There are several **potted plants** placed along the street, adding a touch of greenery to the urban environment. **A traffic light** can be seen in the background, indicating that the taxi is driving through an intersection. The man's unique choice of ironing location and the busy street create an interesting and unconventional scene.



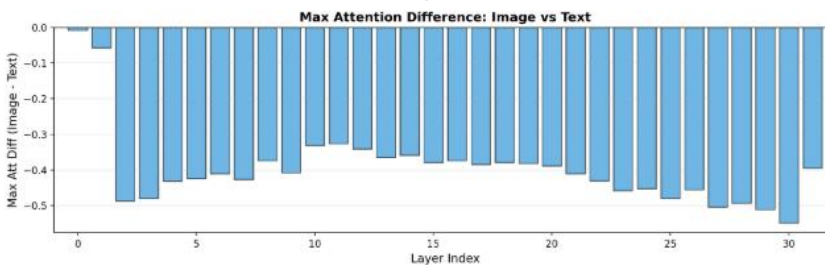
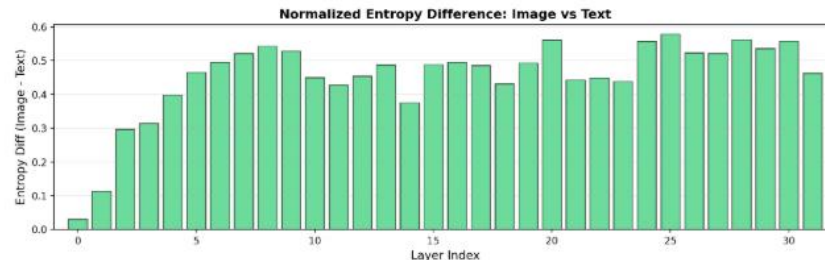
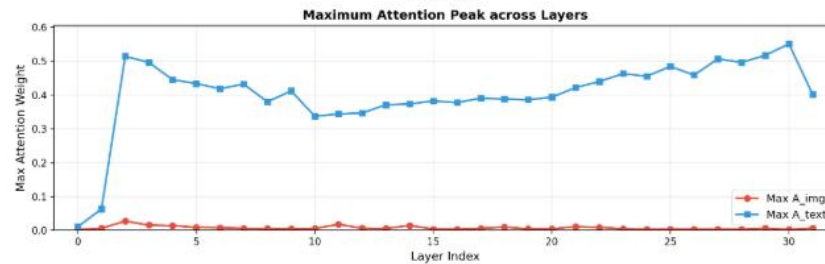
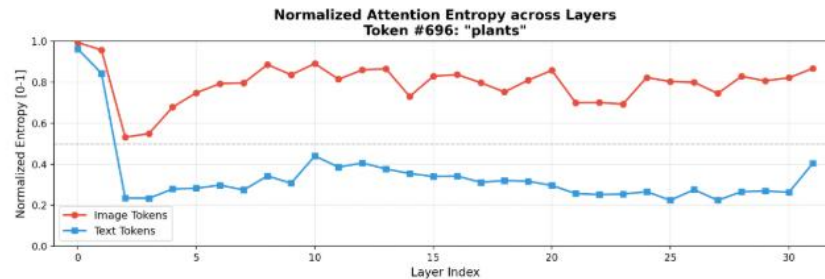
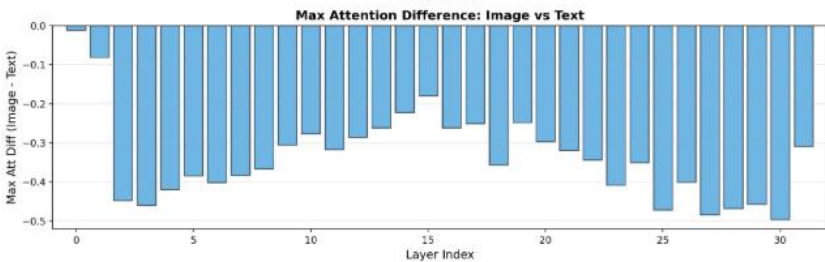
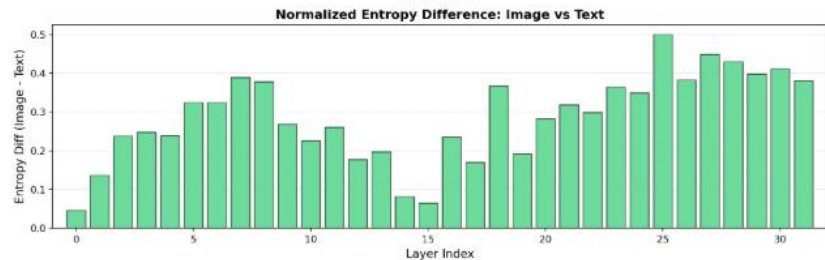
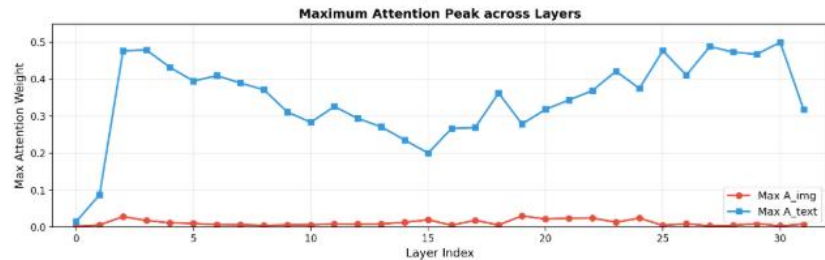
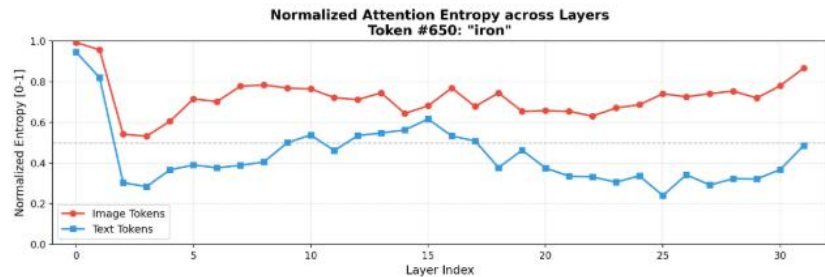
Local visual continuity

This photo shows the shelves inside a refrigerator, neatly arranged with a variety of foods and beverages. Toward the back in the center is a bottle of milk with a red cap; in front of it is a food storage container with a green lid, and on top of that is a package of food that looks like pancakes or flatbread.

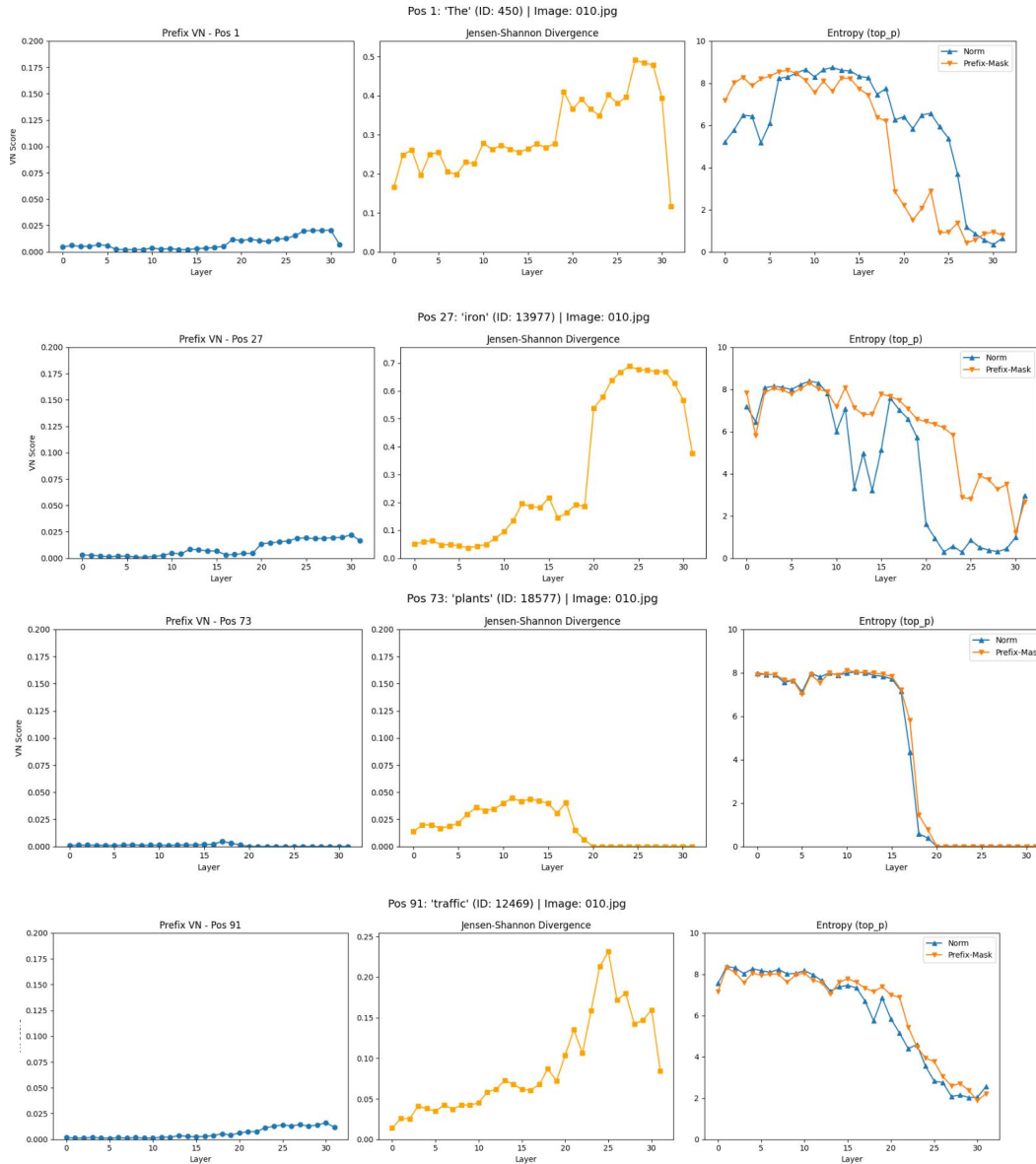
On the left, you can see a bag of orange carrots and a box of fresh strawberries; the strawberries are in a clear plastic container and are brightly colored. Behind the carrots are several stacked takeout or storage containers.

On the right side, a box of blueberries sits at the front, while various yogurts are arranged behind it, including a large tub of yogurt and several individually packaged small cups. Overall, this refrigerator shelf primarily holds fruits, vegetables, and dairy products; the food selection is relatively healthy, and the arrangement is fairly orderly.

Others



必要性VN





Others

必要性: 当前 token 在第 L 层, 视觉路径带来了多少额外增益

$$N_t = \log p(y_t | full) - \log p(y_t | text)$$

充分性: 用 part 相比 text-only 的增益恢复比例

$$S_t = \frac{\log p(y_t | part) - \log p(y_t | text)}{\log p(y_t | full) - \log p(y_t | text) + \epsilon}$$

维护一个全局证据向量 \mathbf{g}_t

在中后层做 logit-bias

$$z_{t+1, \mathcal{I}}^{(\ell, h)} \leftarrow z_{t+1, \mathcal{I}}^{(\ell, h)} + \lambda_t^{(\ell)} b \log(\mathbf{g}_t + \epsilon)$$



南京航空航天大学

Nanjing University of Aeronautics and Astronautics



模式分析与机器智能
工业和信息化部重点实验室

MIT Key Laboratory of
Pattern Analysis & Machine Intelligence

THANKS
