

See What You Are Told: Visual Attention Sink In Large Multimodal Models

Seil Kang, Jinyeong Kim, Junhyeok Kim, Seong Jae Hwang

Yonsei University (ICLR 2025)



Jaemin Kim



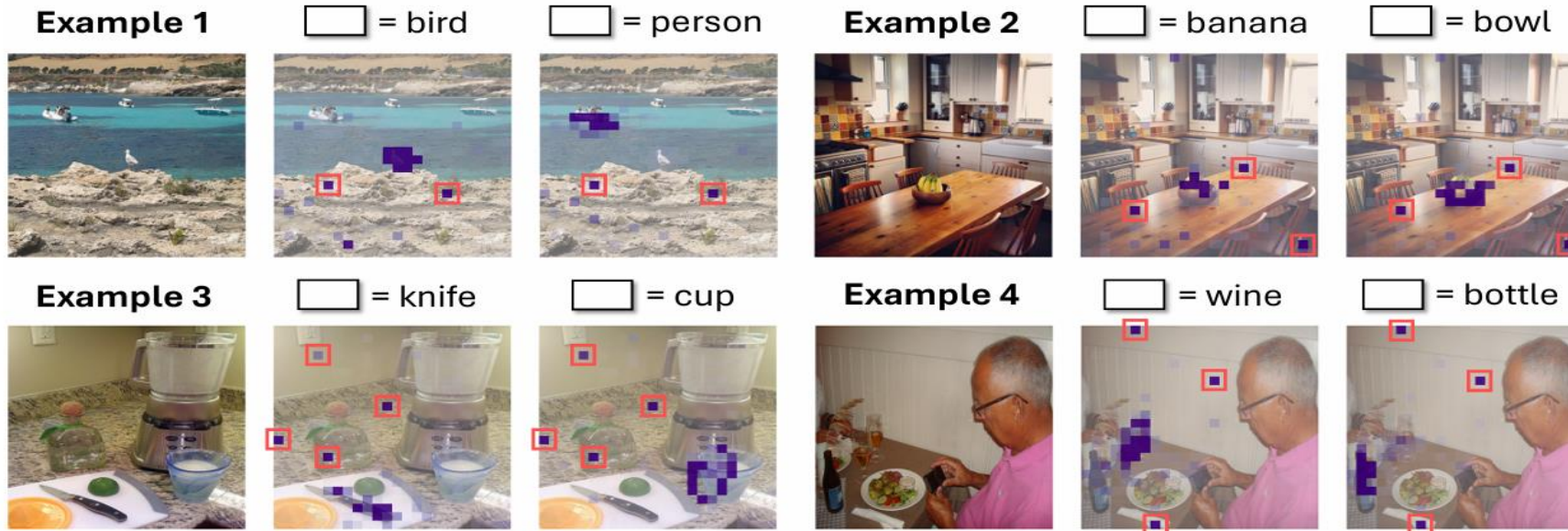
2025-09-10

LMMs see images by leveraging the attention mechanism between text and visual tokens.

LMMs should primarily attend to the visual tokens that are relevant to each text token.

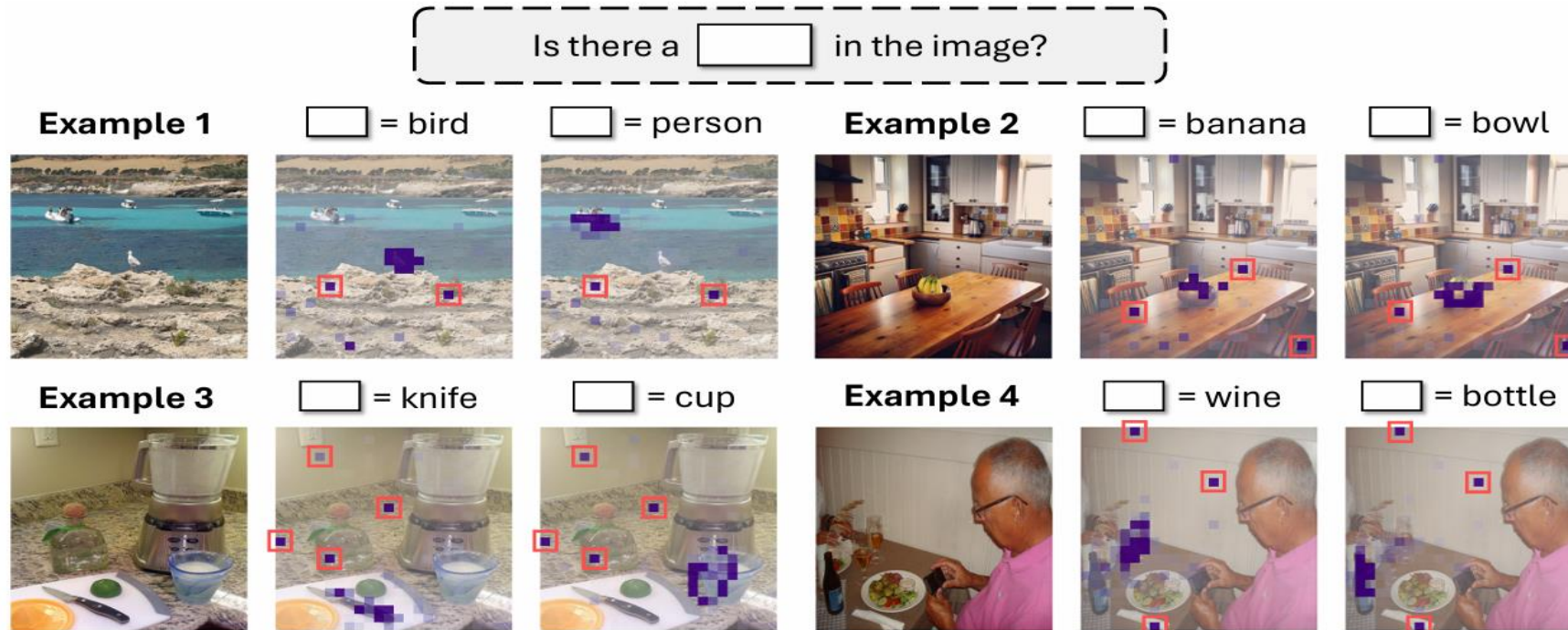
But..

Is there a in the image?



[Visual attention maps of LLaVa-1.5-7B between specified text tokens and visual tokens]

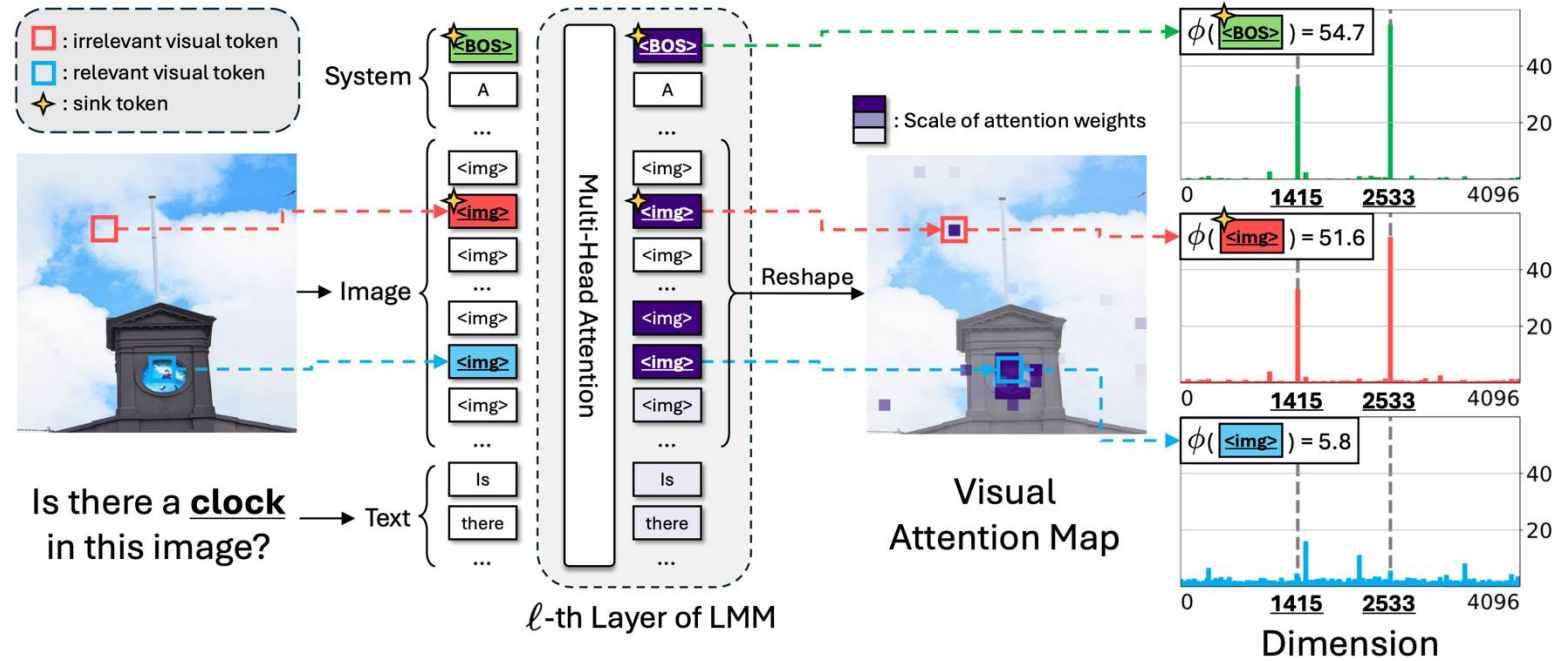
- Uncovered the **underlying properties** of those irrelevant visual tokens
- Found that they are **unnecessary for the model's functioning**
- Proposed **VAR**(Visual Attention Redistribution), a method of recycling surplus attention weights



[Visual attention maps of LLaVa-1.5-7B between specified text tokens and visual tokens]

Model attends to some visual tokens **which are irrelevant to the corresponding text token**, and they exist in **fixed locations**, regardless of the specific text token.

Refer to this phenomenon as **visual attention sink**.



[Illustration of typical architecture of LMMs and investigation of visual attention sink]

$$\phi(x) = \max_{\check{d} \in \mathcal{D}_{\text{sink}}} \left| x[\check{d}] / \sqrt{\frac{1}{D} \sum_{d=1}^D x[d]^2} \right|$$

$\mathcal{D}_{\text{sink}}$ is a set of fixed dimensions, where massive activations are found

[Sink dimension value $\phi(x)$]

With given hidden state $x \in \mathbb{R}^D$, and $x[d]$ is the d -th dimension of the hidden state

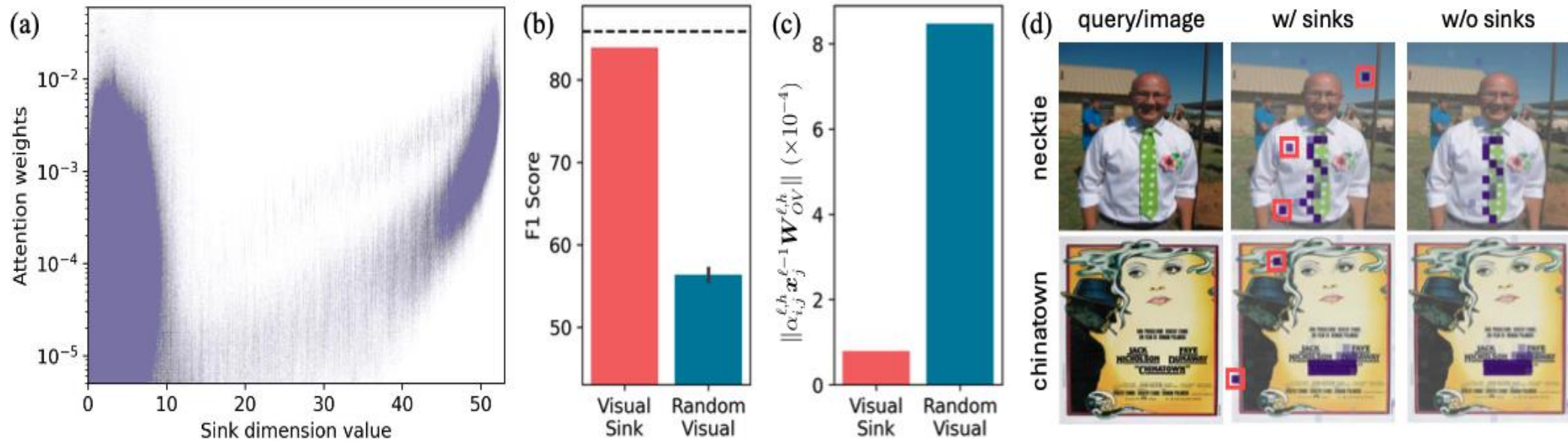
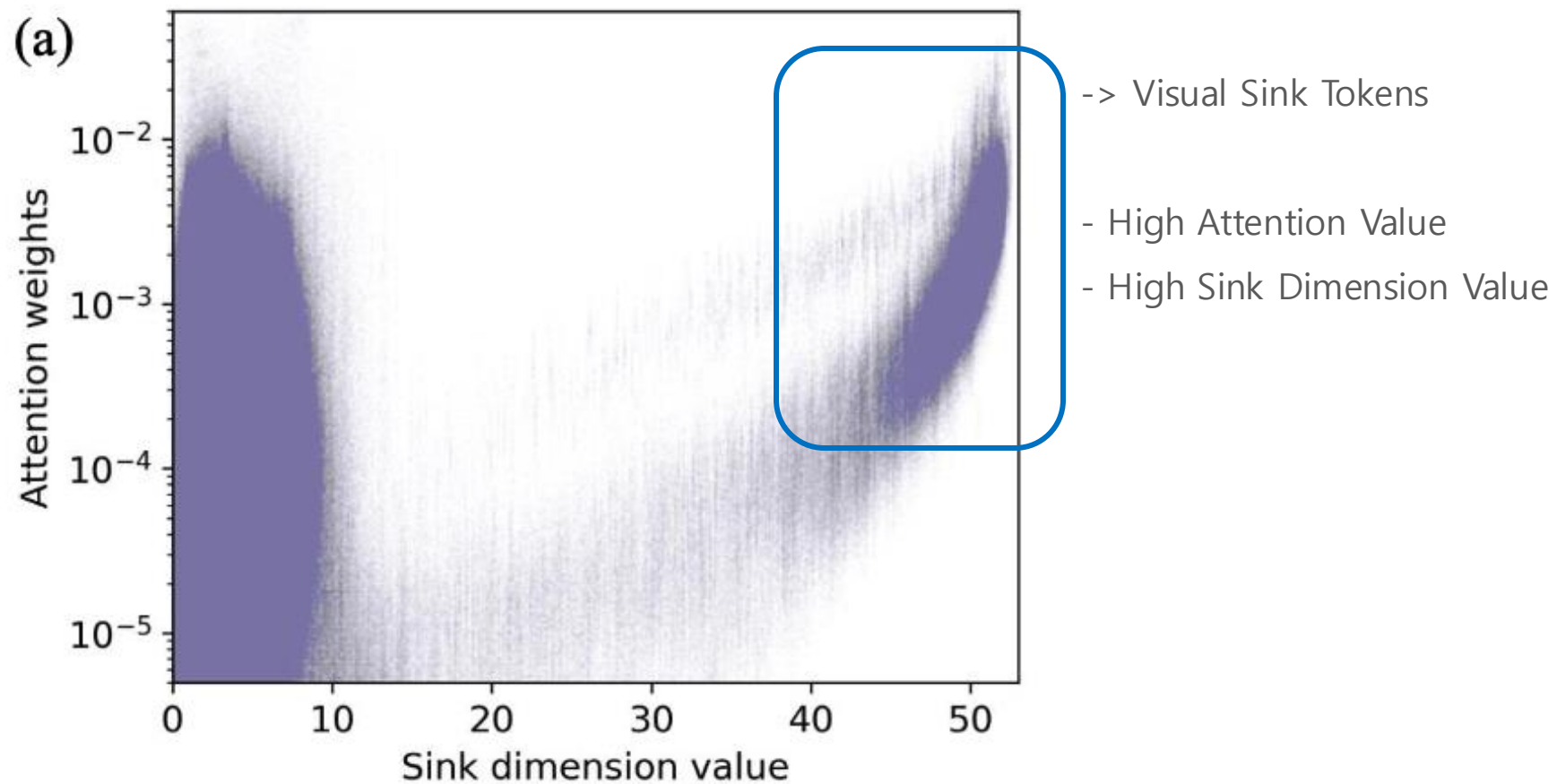
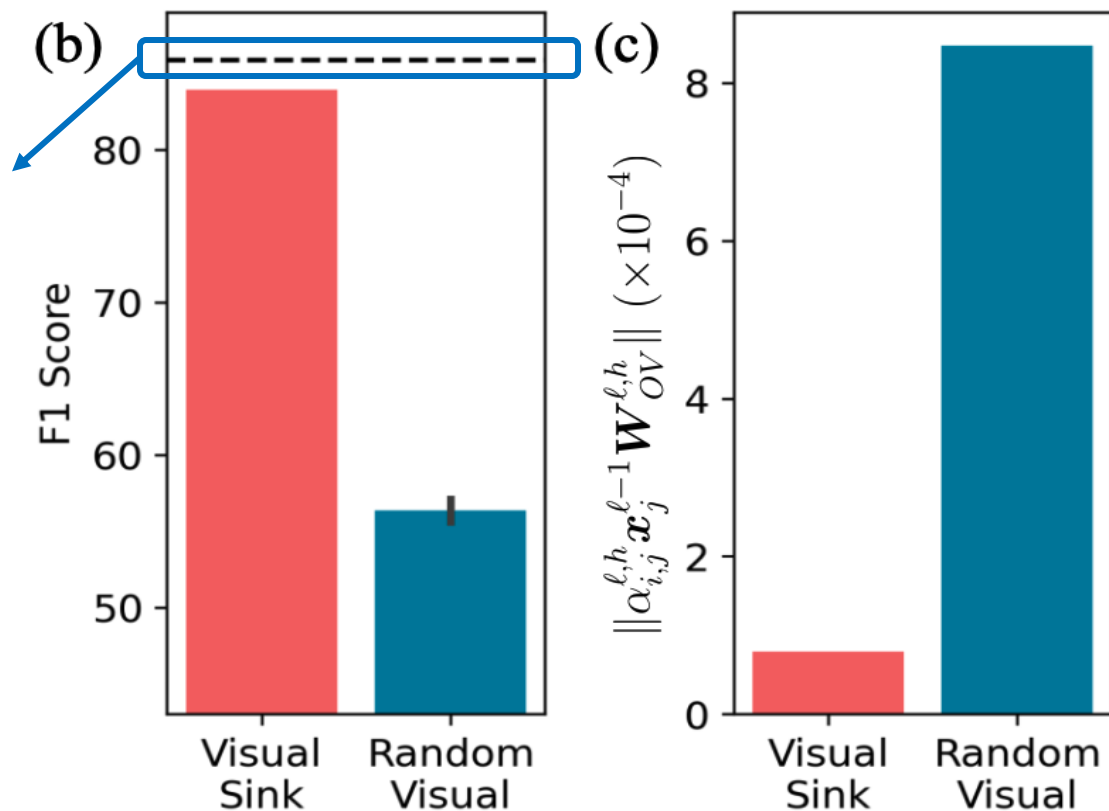


Figure 3: Analysis of visual sink tokens



Scatter plot of sink dimension values and attention weights of visual tokens

Original model performance

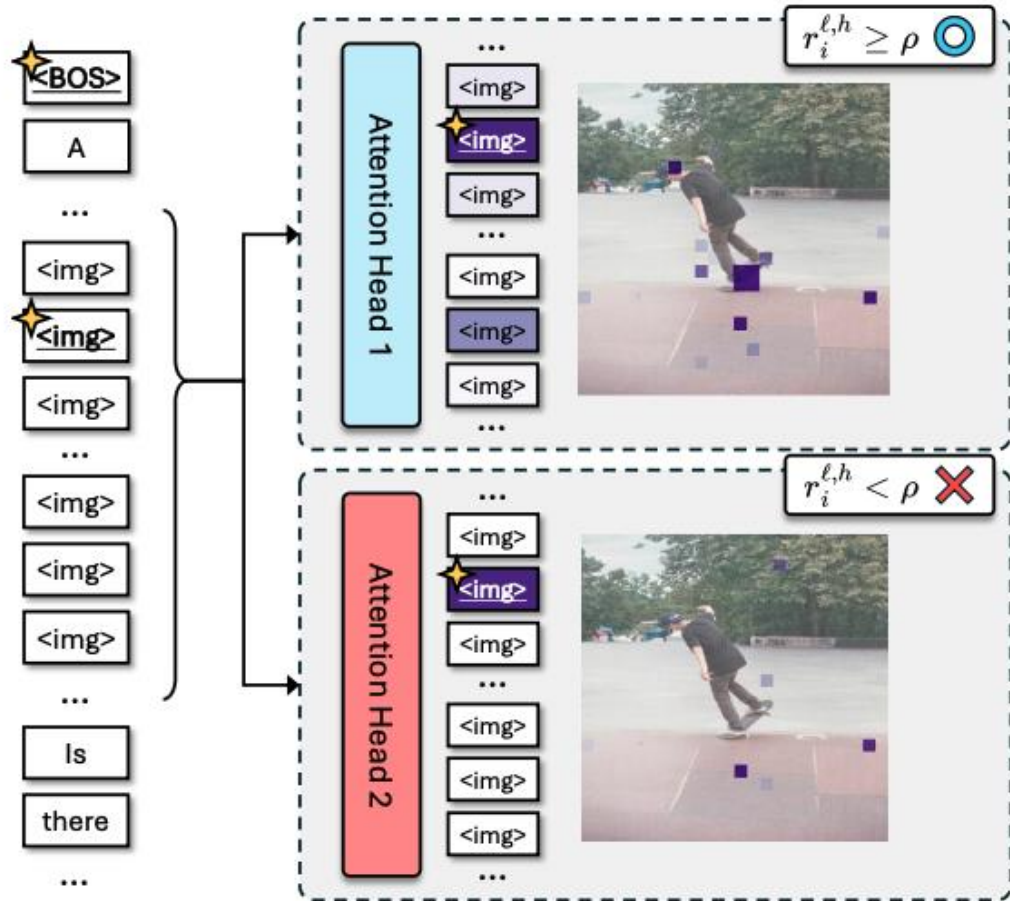


(b) Performance comparison between masking visual sink tokens and masking the same number of random visual tokens

(c) Average attention contributions of visual sink tokens and random visual tokens



Can we recycle surplus attentions in visual attention sink?
They don't even contribute to the model's output!



(a) Select image-centric heads

ρ is a hyperparameter that controls the number of selected heads



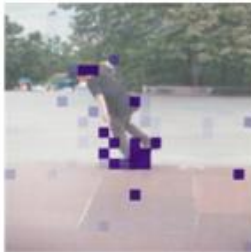







Table 4: Ablation study on selecting image-centric heads.

Setting	POPE		MME	MM-Vet
	F1	Acc.	Score	Score
Baseline	85.9	84.8	1495.5	31.1
w/o Head selection	0.0	0.0	0.0	0.0
w/ Head selection (Ours)	86.5	85.9	1513.8	33.7

$$r_i^{\ell,h} = \frac{\sum_{j \in \mathcal{I}_{\text{vis}} \setminus \check{\mathcal{I}}_{\text{vis}}^{\ell}} \alpha_{i,j}^{\ell,h}}{\sum_{j \in \mathcal{I}_{\text{vis}}} \alpha_{i,j}^{\ell,h}}$$

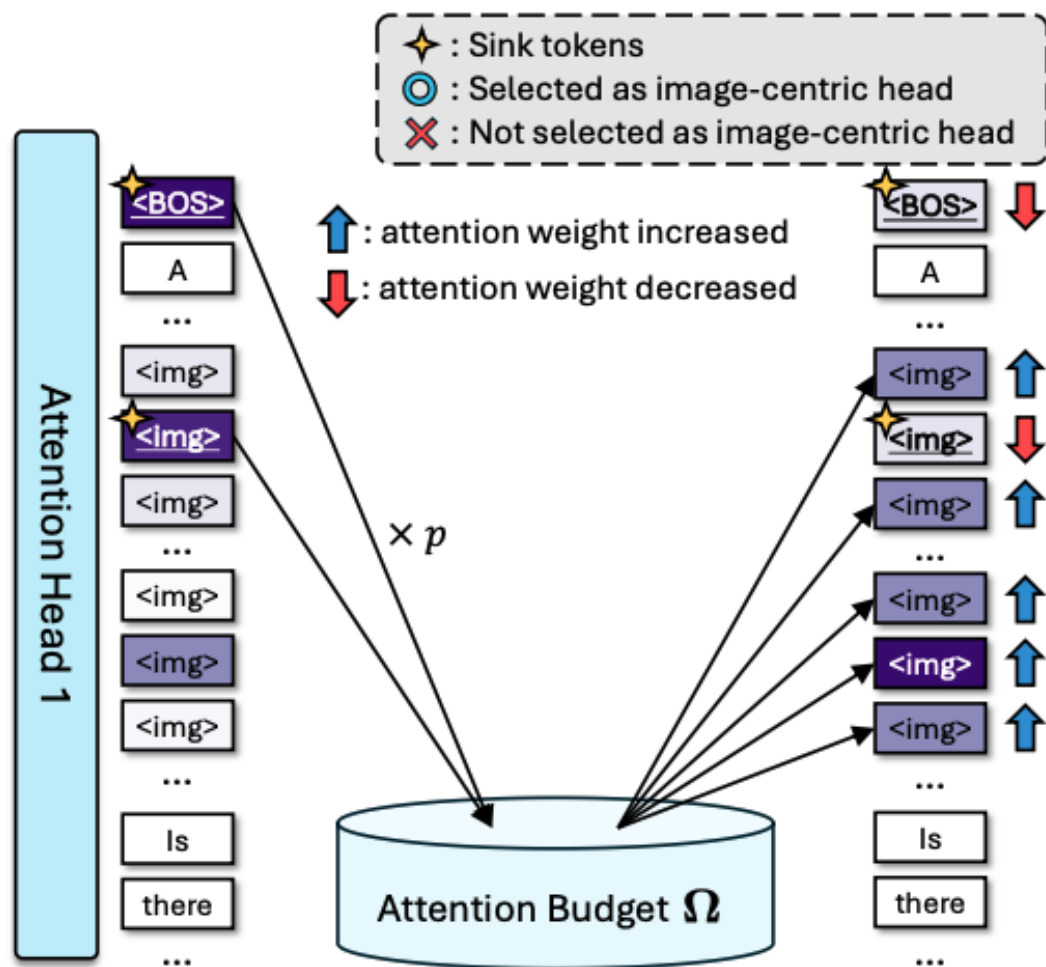
where \mathcal{I}_{vis} and $\mathcal{I}_{\text{vis}} \setminus \check{\mathcal{I}}_{\text{vis}}^{\ell}$ denotes the set of all visual tokens and the visual non-sink tokens

$\alpha_{i,j}^{\ell,h}$ is an attention weight with ℓ indicating layer, h indicating attention head, i indicating query token, and j indicating key/value token.

	Image and Questions		Heads w/ high visual non-sink ratio		Heads w/ low visual non-sink ratio	
Example 1		Is there a skateboard in this image?				
Example 2		Is there a red coat in the image?				

[Visualization of the attention heads sorted by visual non-sink ratio]

The attention heads with high visual non-sink ratio are selected as the image-centric heads



(b) Redistribute attention weights

p ($0 \leq p \leq 1$) is a portion which controls the amount of attention weights to redistribute

Attention weights to the visual non-sink tokens updated as follows:

$$\check{\alpha}_{i,j} = \alpha_{i,j} + \Omega \cdot \frac{\alpha_{i,j}}{\sum_{j \in \mathcal{I}_{\text{vis}} \setminus \check{\mathcal{I}}_{\text{vis}}} \alpha_{i,j}}.$$

Attention weights of the sink tokens decrease to $(1 - p) \cdot \alpha_{i,j}$ and attention budget is calculated as $\Omega = p \cdot \sum_{j \in L} \alpha_{i,j}$ (L is indices of the sink tokens).

* $i \in L_{\text{txt}}$ noting that the redistribution of attention weights is applied to all text tokens.

Base models:

Employed 6 base models (Model lists can be found in Table 1)

Tasks and benchmarks:

(1) *General vision-language tasks* (10 benchmarks listed in the Table 1 at next slide)

assesses comprehensive multimodal capabilities of LMMs

(2) *Visual Hallucination tasks* (CHAIR, POPE, MMHal-Bench)

evaluates whether the response of the model is consistent with the image content

to

ensure the trustworthiness and reliability of the model

(3) *Vision-centric tasks* (MMVP, CV-Bench2D, CV-Bench3D)

evaluates visual understanding capabilities, such as determining the spatial relationship between objects in the image

Table 1: Benchmark results on general vision-language task.

Model	VQA ^{v2}	GQA	VizWiz	SQA ^I	VQA ^T	MME	MMB ^{en}	SEED ^I	LLaVA ^W	MM-Vet
LLaVA-1.5-7B	78.5	62.0	50.0	66.8	58.2	1495.5	64.3	58.6	65.4	31.1
+ Ours	78.6	63.5	53.7	67.3	58.6	1513.8	65.1	60.7	68.1	33.7
LLaVA-1.5-13B	80.0	63.3	53.6	71.6	61.3	1501.2	67.7	61.6	72.5	36.1
+ Ours	81.2	64.9	57.2	72.2	62.1	1534.3	68.1	62.3	74.1	38.4
LLaVA-1.5-HD-13B	81.8	64.7	57.5	71.0	62.5	1500.1	68.8	62.6	72.0	39.4
+ Ours	82.0	65.1	58.8	71.3	63.0	1505.2	69.5	63.3	73.8	40.6
VILA-13B	80.8	63.3	60.6	73.7	66.6	1507.1	70.3	62.8	73.0	38.8
+ Ours	81.2	63.6	64.2	74.7	67.3	1512.7	71.7	63.0	75.7	39.7
Qwen2-VL-7B	82.5	64.5	65.4	74.1	84.3	1672.3	83.0	77.9	75.6	63.2
+ Ours	82.8	64.7	67.7	74.2	84.9	1688.5	83.3	78.1	77.3	63.5
InternVL2-8B	82.0	63.2	63.0	74.2	77.3	1648.1	81.7	76.2	73.2	60.0
+ Ours	82.5	63.5	65.1	74.7	78.0	1655.4	82.3	77.1	75.1	61.2

Table 2: Benchmark results on visual hallucination task.

Model	CHAIR		POPE (all)		MMHal	
	$C_S \downarrow$	$C_I \downarrow$	F1 \uparrow	Acc. \uparrow	Score \uparrow	Hall. \downarrow
LLaVA-1.5-7B	45.0	14.7	85.90	84.76	2.36	51.0
+ Ours	43.2	13.8	86.53	85.87	4.26	45.1
LLaVA-1.5-13B	20.6	6.2	85.90	85.47	2.42	44.3
+ Ours	17.3	5.1	86.12	86.58	4.38	42.7
LLaVA-1.5-HD-13B	42.9	13.2	87.1	85.0	2.35	43.7
+ Ours	40.6	12.8	87.7	87.9	4.15	43.1
VILA-13B	31.0	8.8	84.2	83.58	2.40	44.7
+ Ours	29.7	8.0	85.1	85.4	4.19	42.9
Qwen2-VL-7B	30.5	8.4	87.0	87.5	2.41	44.1
+ Ours	30.1	8.2	87.4	88.2	4.09	43.5
InternVL2-8B	32.4	9.7	86.9	87.8	2.45	43.9
+ Ours	31.8	9.3	87.5	88.6	4.17	42.7

Table 3: Benchmark results on vision-centric task.

Model	MMVP	CV-Bench ^{2D}	CV-Bench ^{3D}
LLaVA-1.5-7B	3.33	56.8	58.4
+ Ours	9.33	57.6	59.0
LLaVA-1.5-13B	24.7	58.2	58.4
+ Ours	28.0	59.6	59.9
LLaVA-1.5-HD-13B	36.0	62.7	65.7
+ Ours	39.1	63.8	66.8
VILA-13B	23.1	58.6	57.9
+ Ours	28.6	59.7	59.8
Qwen2-VL-7B	52.1	63.5	67.6
+ Ours	55.6	63.6	67.7
InternVL2-8B	51.3	61.8	66.8
+ Ours	56.7	62.1	67.1

[Conclusion]

_Phenomenon of Visual Attention Sink

_Investigated their inherent properties

_Recycling their surplus attention weights by VAR(Visual Attention Redistribution) method

_VAR improves the performance

Thank You
