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

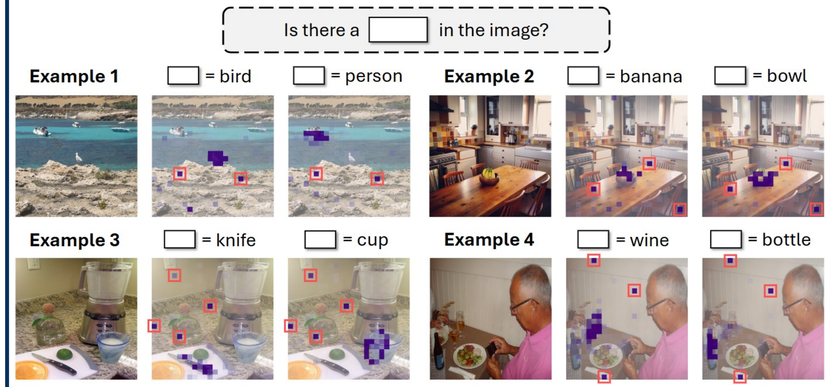
Seil Kang* Jinyeong Kim* Junhyeok Kim Seong Jae Hwang† (* equal contribution)
{seil, jinyeong1324, timespt, seongjae}@yonsei.ac.kr

ICLR International Conference On Learning Representations

YONSEI UNIVERSITY

A. Motivation

LVLMs consistently give attention to irrelevant visual tokens.

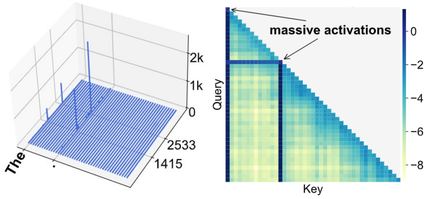


Q: ① *Why* does this phenomenon occur?
② *How* does this phenomenon affect the performance?

B. Preliminaries

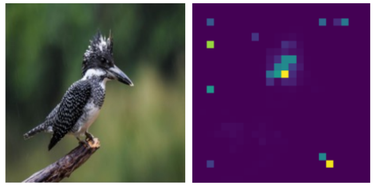
***"Attention sink"* has been observed in LLMs and ViTs.**

LLMs



[1] Sun et al., Massive Activations in Large Language Models

ViTs

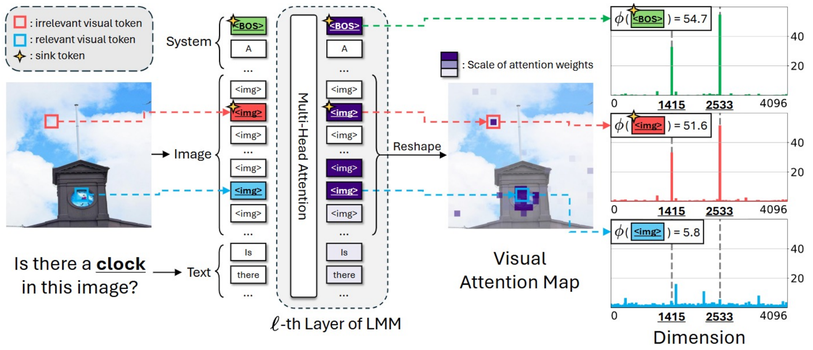


[2] Darcet et al., Vision Transformers Need Registers

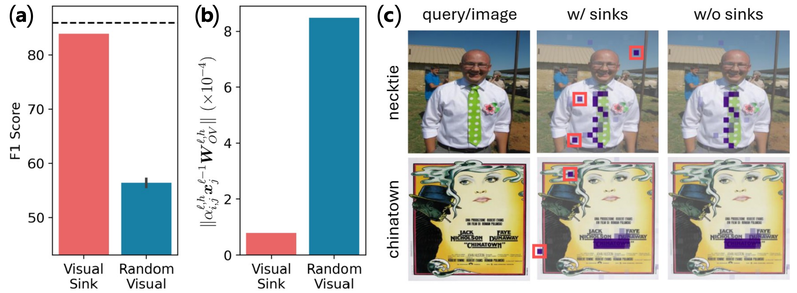
Attention sink occurs...
① by *massive activation* in specific dims (e.g., 1415/2533 @ LLaMA-2-7B)
② at *less meaningful* tokens (e.g., <BOS>, ".", ":" in LLMs / background in ViTs)

C. Analysis

① Irrelevant visual tokens exhibit *massive activation*.
Let's call these tokens "visual sink tokens".



② Visual sink tokens are *less meaningful*.



(a) Masking visual sink tokens has little impact on performance.
(b) Visual sink tokens have extremely low attention contributions.
(c) Visual sink tokens are mostly located in the background.

Can we recycle surplus attention in visual attention sink?
Attention weights in sink tokens = ***FREE "attention budget"***

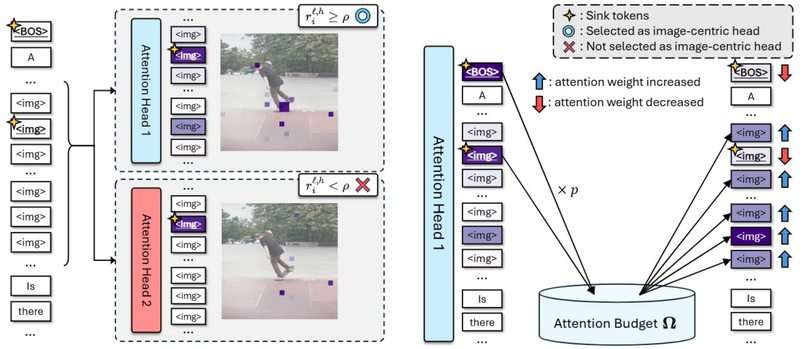
D. Application

See What You Are Told!
Visual Attention Redistribution (VAR)

See you again @ *CVPR2025*

➢ **Step 1. Select Image-centric heads**
Not all attention heads focus on image tokens. Heads with high visual non-sink attention are considered image-centric.

➢ **Step 2. Redistribute attention weights**
Attention weights: sink tokens → visual non-sink tokens



E. Results

***Understanding Attention mechanism* improves LVLMs *for FREE*!**

Model	VQA^{v2}	GQA	VizWiz	SQA^{I}	VQA^{T}	MME	CHAIR		POPE (all)		MMVP	CV-Bench^{2D}
LLaVA-1.5-7B	78.5	62.0	50.0	66.8	58.2	1495.5	45.0	14.7	85.90	84.76	3.33	56.8
+ Ours	**78.6**	**63.5**	**53.7**	**67.3**	**58.6**	**1513.8**	**43.2**	**13.8**	**86.53**	**85.87**	**9.33**	**57.6**
LLaVA-1.5-13B	80.0	63.3	53.6	71.6	61.3	1501.2	20.6	6.2	85.90	85.47	24.7	58.2
+ Ours	**81.2**	**64.9**	**57.2**	**72.2**	**62.1**	**1534.3**	**17.3**	**5.1**	**86.12**	**86.58**	**28.0**	**59.6**
LLaVA-1.5-HD-13B	81.8	64.7	57.5	71.0	62.5	1500.1	42.9	13.2	87.1	85.0	36.0	62.7
+ Ours	**82.0**	**65.1**	**58.8**	**71.3**	**63.0**	**1505.2**	**40.6**	**12.8**	**87.7**	**87.9**	**39.1**	**63.8**
VILA-13B	80.8	63.3	60.6	73.7	66.6	1507.1	31.0	8.8	84.2	83.58	23.1	58.6
+ Ours	**81.2**	**63.6**	**64.2**	**74.7**	**67.3**	**1512.7**	**29.7**	**8.0**	**85.1**	**85.4**	**28.6**	**59.7**
Qwen2-VL-7B	82.5	64.5	65.4	74.1	84.3	1672.3	30.5	8.4	87.0	87.5	52.1	63.5
+ Ours	**82.8**	**64.7**	**67.7**	**74.2**	**84.9**	**1688.5**	**30.1**	**8.2**	**87.4**	**88.2**	**55.6**	**63.6**
InternVL2-8B	82.0	63.2	63.0	74.2	77.3	1648.1	32.4	9.7	86.9	87.8	51.3	61.8
+ Ours	**82.5**	**63.5**	**65.1**	**74.7**	**78.0**	**1655.4**	**31.8**	**9.3**	**87.5**	**88.6**	**56.7**	**62.1**