THE ANCHORING EFFECT OF ATTENTION SINKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Attention Sinks (AS), tokens that accumulate disproportionately high attention weights, are commonly regarded as detrimental artifacts in Large Vision-Language Models (LVLMs), often associated with hallucination and attention redundancy. In this work, we show that this view is incomplete: AS also serve as structural anchors in LVLMs. Specifically, AS persist across layers and attention heads, while targeted perturbations induce disproportionate changes in attention organization and representation geometry. Further masking experiments reveal their dual role: Masking sink attention can mitigate hallucinations, but this benefit may come at the cost of degraded visual reasoning and broader model capabilities. Motivated by these findings, we propose Gated Sink Attention Redistribution (G-SAR), a simple input-adaptive gating mechanism that learns how much attention to release from sink tokens and redistributes the released mass toward visual non-sink tokens. Across three LVLMs and diverse general visual understanding, hallucination, and visual reasoning benchmarks, G-SAR consistently improves overall performance while reducing hallucination. Our results suggest that AS should be regulated rather than eliminated, highlighting adaptive sink control as a more principled strategy to balance visual grounding and reasoning in LVLMs.

1 INTRODUCTION

Attention **S**inks (**AS**), a phenomenon in which Large Language Models (LLMs) assign disproportionately high attention weights to a small subset of tokens, have attracted increasing interest in streaming and long-context generation (Xiao et al., 2024; Han et al., 2024; Yang et al., 2025; Gu et al., 2025; Ge et al., 2024), efficient inference (Zhang et al., 2024; Chen et al., 2025b; Wu & Tu, 2024), and hallucination analysis and mitigation (Huang et al., 2024; Zhang et al., 2025b; Wang et al., 2025b; Tang et al., 2025b;a).

Large vision-language models (LVLMs) exhibit AS not only at the beginning of the textual sequence but also among visual tokens produced by the vision encoder (Luo et al., 2025; Kang et al., 2025; Darcet et al., 2024; Tang et al., 2025a; Huang et al., 2024), introducing additional complexity through cross-modal interactions. Existing studies have primarily treated AS in LVLMs as detrimental artifacts associated with attention redundancy (Takezoe et al., 2026; Zhang et al., 2025c; Ye et al., 2025; Kang et al., 2025). Excessive attention concentration on low-information tokens has been associated with inefficient use of visual information (Huang et al., 2024; Tang et al., 2025a) and increased object hallucination (Wang et al., 2025b; Tang et al., 2025b;a; Sun et al., 2026a), motivating methods that suppress or regulate sink attention.

However, this suppression-oriented view leaves an important question unanswered: are AS merely detrimental artifacts, or do they also play a functional role in LVLM computation? If AS contribute to the model's internal organization, suppressing them may disrupt not only excessive attention concentration but also useful computational structure. Understanding this role is therefore essential for determining how AS should be handled, rather than simply whether they should be removed.

In this work, we conduct a systematic analysis of AS in LVLMs and uncover a previously underexplored **anchoring effect**. We show that AS behave as persistent structural anchors: they repeatedly dominate attention across Transformer layers and heads, while targeted perturbations induce disproportionately large changes in the attention organization and representation geometry of unmasked tokens. Importantly, this anchoring role coexists with their connection to hallucination. Targeted masking shows that reducing sink attention can mitigate hallucination, but can impair visual rea-

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

soning and broader model capabilities. These findings suggest that the key issue may not lie in the existence of AS themselves, but in the excessive concentration of attention on them.

Motivated by this observation, we propose **Gated Sink Attention Redistribution (G-SAR)**, a simple, input-adaptive gating mechanism that *adaptive regulates rather than eliminates* sink attention. Conditioned on the multimodal input, G-SAR learns how much attention mass to retain at each sink and releases only the excess for redistribution toward visual non-sink tokens. By conserving the total attention mass and using the pretrained query–visual attention pattern as the routing prior, G-SAR alleviates excessive concentration while retaining the anchoring contribution of AS. Across three representative LVLMs and diverse benchmarks covering general visual understanding, hallucination, and visual reasoning, G-SAR consistently improves overall performance while reducing hallucination. We further examine the impact of different redistribution strategies on G-SAR.

Our main contributions are summarized as follows:

- We uncover an underexplored **anchoring effect** of AS in LVLMs. AS exhibit persistent dominance across layers and heads, and perturbing them disproportionately alters attention organization and representation geometry, indicating that they occupy structurally influential positions in model computation.
- We conduct a systematic empirical analysis of how AS affect downstream LVLM capabilities and reveal their **dual role**: reducing sink attention can mitigate hallucination, but at the cost of impairing visual reasoning and broader model capabilities. This finding motivates a shift from eliminating AS to adaptively regulating their attention strength.
- We introduce **G-SAR**, a simple input-adaptive gating mechanism that learns how much attention to release from AS and redistributes the released mass toward visual non-sink tokens. Across multiple LVLMs and benchmarks, G-SAR reduces hallucination while improving general visual understanding and reasoning. We further systematically investigate how different redistribution strategies affect its performance.

2 Related Work

Large vision-language models. LVLMs connect a pretrained vision encoder to an LLM through a simple cross-modal interface. Flamingo (Alayrac et al., 2022) uses gated cross-attention to fuse frozen vision and language backbones. BLIP-2 and InstructBLIP (Li et al., 2023; Dai et al., 2023) adopt a Q-Former that queries visual features before feeding them into the LLM. LLaVA (Liu et al., 2023) showed that a simple linear projector with visual instruction tuning already performs competitively, and this design now underpins most follow-up work. Recent models push toward fine-grained perception such as OCR-intensive understanding, multilingual interaction, and spatial grounding (Bai et al., 2025; Peng et al., 2024; Wang et al., 2025a), and now power applications in autonomous driving (Sima et al., 2025; Wang, 2024; Chen et al., 2024; 2025a), document understanding (Duan et al., 2025; Wu et al., 2025), and general multimodal assistance (Jia et al., 2025; Shen et al., 2023).

Attention sinks. An attention sink is a token that consistently receives disproportionately large attention mass. In LLMs, attention sinks support long-context generation (Xiao et al., 2024; Yang et al., 2025; Ge et al., 2024), enable efficient inference (Zhang et al., 2024; Chen et al., 2025b; Wu & Tu, 2024), and facilitate pruning (Ma et al., 2023; Sandoval-Segura et al., 2026). Mechanistic studies point to two explanations: AS may emerge early in training to establish stable coordinate systems in high-dimensional representation spaces (Ruscio et al., 2025), or they may arise alongside compression valleys, both driven by large residual stream activations (de Llano et al., 2026; Sun et al., 2026b). In LVLMs, AS have been observed in object hallucination tasks, where certain tokens consumed excessive attention and were termed summary tokens (Huang et al., 2024; Tang et al., 2025a; Sun et al., 2026a). Several methods now redistribute this attention to reduce visual redundancy (Kang et al., 2025; Luo et al., 2025). What remains unclear is whether AS serve a functional role in LVLMs beyond being redundant, and how suppressing them affects downstream tasks. We find that attention sinks act as persistent structural anchors in LVLMs, that removing them creates a trade-off between hallucination and visual reasoning, and that a simple gating mechanism can effectively balance this trade-off.

3 PRELIMINARIES

We use bold lowercase for vectors ($\mathbf{a}$, with a_i as the i-th entry and $\|\mathbf{a}\|$ the Euclidean norm) and bold uppercase for matrices ($\mathbf{A}$, with $\mathbf{A}_i$ as the i-th column and $\|\mathbf{A}\|$ the spectral norm). The i-th standard basis vector is e_i.

Transformer. Let $X = [x_1, \ldots, x_T] \in \mathbb{R}^{d \times T}$ denote a sequence of T textual and visual token embeddings. For a Transformer layer ℓ and attention head h, the query, key, and value representations are given by

$$Q^{\ell,h} = W_Q^{\ell,h} X^{\ell-1}, \qquad K^{\ell,h} = W_K^{\ell,h} X^{\ell-1}, \qquad V^{\ell,h} = W_V^{\ell,h} X^{\ell-1}, \tag{1}$$

where $W_Q^{\ell,h}, W_K^{\ell,h}, W_V^{\ell,h}$ denote the corresponding projection matrices. The attention matrix is computed as

$$A^{\ell,h} = \mathrm{softmax}\left(\frac{(Q^{\ell,h})^\top K^{\ell,h}}{\sqrt{d_h}} + M \right), \tag{2}$$

where d_h is the head dimension and M denotes the causal attention mask. Accordingly, $A_{ij}^{\ell,h}$ represents the attention weight assigned by query token i to key token j at layer ℓ and head h.

The corresponding attention output for token i is

$$u_i^{\ell,h} = \sum_{j=1}^{T} A_{ij}^{\ell,h} v_j^{\ell,h}. \tag{3}$$

Multi-head outputs are subsequently aggregated and processed by the standard residual, normalization, and feed-forward operations. We denote the resulting hidden representation of token i at layer ℓ by h_i^ℓ.

Attention sinks. Following prior work, we define AS as token positions that consistently receive disproportionately high incoming attention (Ruscio et al., 2025; Gu et al., 2025). Given the attention maps from the prefill forward pass, we define the layer- and head-aggregated attention as

$$\bar{A}_{ij} = \frac{1}{LH} \sum_{\ell=1}^{L} \sum_{h=1}^{H} A_{ij}^{\ell,h}, \tag{4}$$

where $\bar{A}_{ij}$ denotes the average attention assigned by query token i to key token j across all layers and attention heads.

Because causal masking allows token j to be attended to only by query positions $i \geq j$, we measure how consistently token j receives high incoming attention over all valid queries. We define its sink frequency as

$$\mathrm{freq}(j) = \frac{1}{N - j + 1} \sum_{i=j}^{N} \mathbf{1}\left\{ \bar{A}_{ij} \geq \tau_{m(j)} \right\}, \tag{5}$$

where N denotes the sequence length, $\mathbf{1}\{\cdot\}$ is the indicator function, $m(j) \in \{\mathrm{txt}, \mathrm{vis}\}$ denotes the modality of the key token j, and $\tau_{m(j)}$ is the corresponding modality-specific adaptive attention threshold. A token is identified as an attention sink if

$$\mathrm{freq}(j) \geq \gamma, \tag{6}$$

where we set $\gamma = 0.5$ in the main experiments.

We refer to sinks at textual positions as **Textual Attention Sinks (TAS)** and those at visual positions as **Visual Attention Sinks (VAS)**. Details of modality-specific threshold construction and qualitative analysis are provided in Appendix A.

4 CHARACTERIZING AND ANALYZING THE ANCHORING EFFECT

Although "anchor" is not uniquely formalized, it broadly refers to a stable reference point that supports internal model computation (Wang et al., 2023; Darcet et al., 2024; Zhang et al., 2025a).

We characterize the anchoring effect of AS through two complementary properties: **persistence**, which measures whether sink behavior is sustained across layers and attention heads, and **influence**, which measures how strongly perturbing AS alters representations and attention organization.

Persistence. Let $\mathcal{S}$ denote the AS identified from the layer- and head-aggregated attention in Sec. 3. For sink $j \in \mathcal{S}$, we measure how persistently it remains a sink across Transformer layers and attention heads. For layer-wise persistence, we average attention over heads,

$$\bar{A}^{\ell}_{ij} = \frac{1}{H} \sum_{h=1}^{H} A^{\ell,h}_{ij}, \tag{7}$$

and compute the layer-specific sink frequency

$$f^{\ell}_{j} = \frac{1}{N-j+1} \sum_{i=j}^{N} \mathbf{1}\left\{ \bar{A}^{\ell}_{ij} \geq \tau_{m(j)} \right\}. \tag{8}$$

The layer-wise persistence of sink j is then

$$P_{\text{layer}}(j) = \frac{1}{L} \sum_{\ell=1}^{L} \mathbf{1}\left\{ f^{\ell}_{j} \geq \gamma \right\}. \tag{9}$$

Similarly, for head-wise persistence, we average attention over layers,

$$\bar{A}^{h}_{ij} = \frac{1}{L} \sum_{\ell=1}^{L} A^{\ell,h}_{ij}, \tag{10}$$

and define

$$f^{h}_{j} = \frac{1}{N-j+1} \sum_{i=j}^{N} \mathbf{1}\left\{ \bar{A}^{h}_{ij} \geq \tau_{m(j)} \right\}, \qquad P_{\text{head}}(j) = \frac{1}{H} \sum_{h=1}^{H} \mathbf{1}\left\{ f^{h}_{j} \geq \gamma \right\}. \tag{11}$$

We report both metrics averaged over TAS and VAS separately.

Influence. We quantify the influence of AS by comparing the original model with a sink-targeted masking intervention. To isolate the propagated effect of sink masking on the remaining representation space, rather than changes directly induced at the intervention targets, all representation- and attention-level comparisons are performed over non-masked positions.

Representation drift. Let $\mathcal{U}$ denote the evaluation token set, e.g., all non-masked tokens or non-masked visual tokens. We quantify the representation change at the final Transformer layer between the original and sink-masked models using

$$D_{\text{L2}} = \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \left\| h^{\text{full}}_{i} - h^{\text{mask}}_{i} \right\|_2, \tag{12}$$

$$D_{\cos} = \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \left[1 - \cos\left(h^{\text{full}}_{i}, h^{\text{mask}}_{i} \right) \right], \tag{13}$$

and

$$D_{\text{rel}} = \frac{1}{|\mathcal{U}|} \sum_{i \in \mathcal{U}} \frac{\left\| h^{\text{full}}_{i} - h^{\text{mask}}_{i} \right\|_2}{\left\| h^{\text{full}}_{i} \right\|_2 + \epsilon}, \tag{14}$$

where h^{full}_{i} and h^{mask}_{i} denote the final-layer representations of token i in the original and sink-masked models, respectively, and ϵ is a small constant (e.g., 10^{-8}) for numerical stability.

Attention distribution shift. For each query i, let $\mathcal{R}_i$ denote the key positions that are both causally valid and non-masked under the intervention. To isolate the redistribution of attention over the remaining support, we restrict the attention distributions of both the original and sink-masked models to $\mathcal{R}_i$ and renormalize:

$$\tilde{A}^{\ell,h,c}_{i,j} = \frac{A^{\ell,h,c}_{i,j}}{\sum_{k \in \mathcal{R}_i} A^{\ell,h,c}_{i,k}}, \qquad j \in \mathcal{R}_i, \quad c \in \{\text{full}, \text{mask}\}. \tag{15}$$

We then quantify the attention distribution shift by computing the KL divergence for each layer, head, and valid query, and averaging across all of them:

$$D_{\mathrm{KL}} = \frac{1}{LH|\mathcal{Q}|} \sum_{\ell=1}^{L} \sum_{h=1}^{H} \sum_{i \in \mathcal{Q}} D_{\mathrm{KL}} \left(\tilde{A}^{\ell,h,\mathrm{full}}_{i,\mathcal{R}_i} \,\middle\|\, \tilde{A}^{\ell,h,\mathrm{mask}}_{i,\mathcal{R}_i} \right), \tag{16}$$

where $\mathcal{Q}$ contains queries with non-empty remaining support.

Table 1: **Persistence of AS across layers and attention heads on MME benchmark.** Subscripts indicate standard deviations across samples, and **Avg.** denotes the average across the three LVLMs.

Model	TAS P_{layer}	TAS P_{head}	VAS P_{layer}	VAS P_{head}
Qwen3-VL-8B	$0.63_{\pm 0.03}$	$0.89_{\pm 0.04}$	$0.51_{\pm 0.02}$	$0.76_{\pm 0.04}$
InternVL3.5-8B	$0.65_{\pm 0.01}$	$0.88_{\pm 0.02}$	$0.53_{\pm 0.04}$	$0.91_{\pm 0.05}$
LLaVA-1.5-7B	$0.92_{\pm 0.00}$	$1.00_{\pm 0.00}$	$0.75_{\pm 0.08}$	$0.91_{\pm 0.06}$
Avg.	$\mathbf{0.73}_{\pm 0.01}$	$\mathbf{0.92}_{\pm 0.02}$	$\mathbf{0.60}_{\pm 0.05}$	$\mathbf{0.86}_{\pm 0.05}$

Table 2: **Influence of AS on representation geometry and attention organization in LLaVA-1.5-7B on MME benchmark.** Sink-targeted masking is compared with random and position-matched controls. Parentheses report ratios to the corresponding random control.

	Textual Attention Sinks			Visual Attention Sinks		
Influence Metric	**Random**	**Nearby**	**Sink**	**Random**	**Nearby**	**Sink**
Representation D_{rel}	0.050	0.067	**0.172** (3.4×)	0.010	0.009	**0.039** (3.9×)
Representation D_{L_2}	6.9	8.6	**22.7** (3.3×)	1.4	1.1	**5.2** (3.7×)
Representation $D_{\cos}$	0.0045	0.0051	**0.0171** (3.8×)	0.0003	0.0003	**0.0012** (4.0×)
Attention D_{KL}	0.0119	0.0029	**1.747** (147×)	0.0002	0.0002	**0.0003** (1.5×)

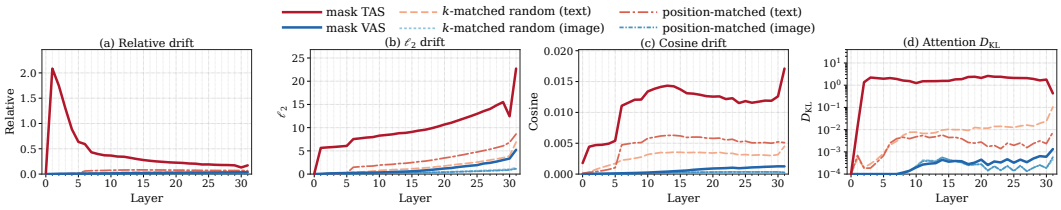

Figure 1: **Layer-wise propagation of sink-induced perturbations in LLaVA-1.5-7B on MME.**

4.1 Persistence of AS Dominance across Layers and Heads

Table 1 evaluates the persistence of textual and visual attention sinks across Transformer layers and attention heads in three representative LVLMs. Across all models, both TAS and VAS exhibit substantial persistence, indicating that sink dominance is not confined to isolated computational components. Specifically, TAS show consistently strong stability, with average P_{layer} and P_{head} of 0.73 and 0.92, respectively. Their head-wise persistence ranges from 0.88 to 1.00, suggesting that textual sinks are stably maintained across attention heads. VAS are also persistent, with average P_{layer} and P_{head} of 0.60 and 0.86, respectively, but exhibit greater model dependence, with layer-wise and head-wise persistence ranging from 0.51–0.75 and 0.76–0.91, respectively.

4.2 AS Preserve Attention and Representation Geometry

We further investigate whether AS play a structural role in maintaining the model's internal organization. We conduct controlled masking experiments in LLaVA-1.5-7B on the full MME benchmark, comparing sink-targeted masking with both modality-matched random masking and position-matched controls. Since AS lie in the extreme high-attention tail, comparable non-sink tokens are

Table 3: **Impact of masking AS on general visual understanding, hallucination, and visual reasoning.** The no-masking baseline (*None*) is highlighted in yellow. Superscripts show changes from the baseline; positive values indicate improvements, including lower CHAIR-S/I scores.

Masked Tokens	General Visual Tasks						Hallucination Tasks					Visual Reasoning Tasks	
	GQA	TextVQA	MMMU	MME_P	MME_C	MM-Vet	POPE-R	POPE-P	POPE-A	CHAIR-S ↓	CHAIR-I ↓	SciQA	MathVista
Qwen3-VL-8B													
None	72.4	82.7	29.0	1716.7	613.9	66.9	92.7	90.1	90.5	24.3	6.8	68.0	37.4
TAS	$73.2_{+0.8}$	$82.8_{+0.1}$	$33.4_{+4.4}$	$1716.7_{0.0}$	$623.1_{+9.2}$	$68.0_{+1.1}$	$92.7_{0.0}$	$91.1_{+1.0}$	$90.9_{+0.4}$	$24.0_{+0.3}$	$6.5_{+0.3}$	$75.0_{+7.0}$	$37.6_{+0.2}$
VAS	$71.0_{-1.4}$	$82.9_{+0.2}$	$28.4_{-0.6}$	$1716.7_{0.0}$	$616.7_{+2.8}$	$64.2_{-2.7}$	$92.8_{+0.1}$	$91.2_{+1.1}$	$90.5_{0.0}$	$22.9_{+1.4}$	$5.9_{+0.9}$	$59.6_{-8.4}$	$35.8_{-1.6}$
TAS+VAS	$0.9_{-71.5}$	$21.2_{-61.5}$	$7.8_{-21.2}$	$659.5_{-1057.2}$	$150.8_{-463.1}$	$3.1_{-63.8}$	$52.1_{-40.6}$	$48.4_{-41.7}$	$50.9_{-39.6}$	$0.6_{+23.7}$	$63.0_{-56.2}$	$1.9_{-66.1}$	$1.7_{-35.7}$
InternVL-3.5-8B													
None	71.0	77.1	53.0	1636.1	622.2	62.9	89.3	87.6	87.6	25.1	7.7	95.2	40.0
TAS	$73.2_{+2.2}$	$77.9_{+0.8}$	$54.2_{+1.2}$	$1525.0_{-111.1}$	$588.9_{-33.3}$	$63.0_{+0.1}$	$89.4_{+0.1}$	$88.2_{+0.6}$	$87.6_{0.0}$	$23.2_{+1.9}$	$7.2_{+0.5}$	$93.8_{-1.4}$	$37.2_{-2.8}$
VAS	$71.6_{+0.6}$	$77.3_{+0.2}$	$53.2_{+0.2}$	$1625.0_{-11.1}$	$602.8_{-19.4}$	$61.5_{-1.4}$	$89.3_{0.0}$	$87.8_{+0.2}$	$87.9_{+0.3}$	$24.8_{+0.3}$	$7.5_{+0.2}$	$95.0_{-0.2}$	$39.6_{-0.4}$
TAS+VAS	$1.0_{-70.0}$	$22.9_{-54.2}$	$5.1_{-47.9}$	$324.8_{-1311.3}$	$42.8_{-579.4}$	$1.0_{-61.9}$	$48.4_{-40.9}$	$50.1_{-37.5}$	$52.7_{-34.9}$	$0.8_{+24.3}$	$65.0_{-57.3}$	$9.4_{-85.8}$	$0.2_{-39.8}$
LLaVA-1.5-7B													
None	61.9	58.2	27.3	1466.9	282.9	31.0	88.3	87.0	84.7	51.0	14.3	64.8	24.4
TAS	$62.1_{+0.2}$	$57.4_{-0.8}$	$26.9_{-0.4}$	$1516.9_{+50.0}$	$305.1_{+22.2}$	$31.2_{+0.2}$	$89.2_{+0.9}$	$87.8_{+0.8}$	$85.4_{+0.7}$	$49.1_{+1.9}$	$13.8_{+0.5}$	$64.8_{0.0}$	$24.1_{-0.3}$
VAS	$62.5_{+0.6}$	$58.0_{-0.2}$	$26.1_{-1.2}$	$1436.4_{-30.5}$	$288.5_{+5.6}$	$31.0_{0.0}$	$88.3_{0.0}$	$87.0_{0.0}$	$84.7_{0.0}$	$50.3_{+0.7}$	$13.9_{+0.4}$	$64.0_{-0.8}$	$23.8_{-0.6}$
TAS+VAS	$1.1_{-60.8}$	$20.0_{-38.2}$	$2.0_{-25.3}$	$662.5_{-804.4}$	$20.8_{-262.1}$	$2.0_{-29.0}$	$45.5_{-42.8}$	$50.5_{-36.5}$	$55.0_{-29.7}$	$4.8_{+46.2}$	$55.0_{-40.7}$	$7.1_{-57.7}$	$1.3_{-23.1}$

typically unavailable for exact attention matching. For each sink type, the same number of tokens is masked within the corresponding modality, and random masking is repeated 20 times and averaged to reduce sampling variance.

AS exert non-local influence on representation geometry. As shown in Table 2, masking TAS or VAS causes substantially greater representation drift than random and position-matched controls. At the final layer, TAS masking yields $3.4\times$–$3.8\times$ larger drift than random-text masking across D_{rel}, D_{L_2}, and $D_{\cos}$, while VAS masking reaches $3.7\times$–$4.0\times$ the random-visual controls. The consistently smaller position-matched effects further rule out generic masking and positional factors, indicating that sink positions disproportionately shape the representations of unmasked tokens.

Sink perturbations exhibit distinct layer-wise dynamics. Figure 1 shows that the structural influence of AS extends throughout the Transformer hierarchy. TAS masking causes a sharp early-layer representational deviation that remains well above random and position-matched controls, with particularly large attention divergence across most layers. In contrast, VAS perturbations are weaker initially but grow with depth, becoming more evident in representation drift while causing limited attention-distribution changes. These contrasting dynamics suggest that textual and visual sinks exert distinct structural influences on internal computation.

TAS predominantly anchor attention organization, whereas VAS primarily support representation geometry. The clearest asymmetry emerges in attention divergence. TAS masking increases D_{KL} from 0.0119 under random-text masking to 1.747, an approximately $147\times$ increase, while the position-matched control remains only 0.0029. In contrast, VAS masking yields only a modest increase in attention divergence (0.0003 versus 0.0002 under random visual masking), despite producing 3.7–$4.0\times$ larger representation drift than the random controls. Together, these results reveal a modality-dependent anchoring structure: TAS exert particularly strong influence on global attention organization, whereas VAS primarily affect representation geometry. Thus, AS are not merely high-attention outliers; they occupy structurally influential positions whose perturbation propagates beyond the sink tokens and reorganizes subsequent model computation.

4.3 Suppressing AS Reveals a Hallucination–Reasoning Trade-off

Motivation. Prior studies (Wang et al., 2025b; Tang et al., 2025b;a) associate excessive attention concentration with visual hallucination and motivate suppressing or redistributing attention sinks. However, Sec. 4.2 reveals that AS also serve as structural anchors that preserve representation geometry and attention organization. This dual role raises a fundamental question: *while suppressing AS may reduce hallucination, does it also disrupt their structural role and thereby impair other visual capabilities, particularly reasoning?*

Experimental settings. We mask TAS, VAS, or both in Qwen3-VL-8B, InternVL3.5-8B, and LLaVA-1.5-7B. We evaluate general visual understanding (GQA, TextVQA, MMMU, MME, and MM-Vet), hallucination (POPE and CHAIR), and visual reasoning (ScienceQA and MathVista). Detailed benchmark and metric descriptions are provided in Sec. 6.1 and Appendix B.

Results and Observations. Table 3 reveals two consistent patterns. **(1) Suppressing AS exposes a hallucination–reasoning trade-off, with modality-dependent effects.** Masking VAS consistently improves or preserves hallucination metrics while degrading visual reasoning across all three models. For example, on Qwen3-VL-8B, CHAIR-S/I decrease from $24.3/6.8$ to $22.9/5.9$, whereas ScienceQA and MathVista drop from $68.0/37.4$ to $59.6/35.8$. TAS show a similar tendency in hallucination mitigation, but their effect on reasoning is more architecture-dependent: ScienceQA improves on Qwen3-VL-8B, decreases on InternVL3.5-8B, and remains unchanged on LLaVA-1.5-7B. **(2) Removing all AS causes severe capability collapse.** Jointly masking TAS and VAS leads to drastic degradation across general understanding and reasoning tasks. For instance, Qwen3-VL-8B drops from 72.4 to 0.9 on GQA and from 68.0 to 1.9 on ScienceQA, with similarly severe degradation across the other models. Overall, these results reveal a dual role of AS: reducing sink attention can mitigate hallucination, but excessive suppression can impair visual reasoning and broader model capabilities. This motivates a more appropriate strategy for AS regulation.

5 G-SAR

Our analysis in Secs. 4.1–4.3 reveals a dual role of AS: excessive sink concentration is associated with hallucination, while AS also serve as structural anchors. Direct suppression of AS therefore reveals a hallucination–reasoning trade-off. Motivated by this, we propose **Gated Sink Attention Redistribution (G-SAR)**, a simple, input-adaptive gating mechanism that learns how much attention to release from AS and redistributes the released mass toward visual non-sink tokens.

5.1 Input-Adaptive Sink Gating

Let $\mathcal{S} = \{s_1, \ldots, s_K\}$ denote the detected AS positions. For each sink s_m at layer ℓ and head h, G-SAR predicts an input-dependent gate conditioned on its key representation $\mathbf{k}_{s_m}^{\ell,h}$ and the mean-pooled query representation $\bar{\mathbf{q}}^{\ell,h}$:

$$g_m^{\ell,h} = \sigma\Big(f_{\ell,h}^{(g)}\left([\mathbf{k}_{s_m}^{\ell,h}; \bar{\mathbf{q}}^{\ell,h}]\right)\Big), \qquad g_m^{\ell,h} \in [0,1], \tag{17}$$

where $f_{\ell,h}^{(g)}$ is a simple MLP, and the mean-pooled query representation $\bar{\mathbf{q}}^{\ell,h}$ is computed exclusively over the visual and prompt tokens, excluding all answer tokens. The gate controls the fraction of attention retained by each sink: $g_m^{\ell,h} \approx 1$ preserves its original attention, whereas a smaller value releases more attention for redistribution. The same gate is shared across all causally valid queries attending to the corresponding sink.

5.2 Sink-to-Visual Attention Redistribution

Let $A_{i,j}^{\ell,h}$ denote the attention assigned by query i to key token j. For each query, G-SAR redistributes released sink attention only to causally valid visual non-sink tokens.

After gating, the total attention mass released from AS is

$$\Delta_i^{\ell,h} = \sum_{m=1}^{K} (1 - g_m^{\ell,h}) A_{i,s_m}^{\ell,h}. \tag{18}$$

We distribute this mass according to the original query–visual attention among the eligible visual non-sink tokens:

$$r_{i,j}^{\ell,h} = \frac{A_{i,j}^{\ell,h}}{\sum_{v \in \mathcal{V}_i^-} A_{i,v}^{\ell,h}}, \qquad j \in \mathcal{V}_i^-, \tag{19}$$

where $\mathcal{V}_i^-$ denotes the causally valid visual non-sink tokens for query i. If $\mathcal{V}_i^-$ is empty, we skip redistribution and retain the original attention distribution for that query. The resulting attention is

$$\widetilde{A}_{i,s_m}^{\ell,h} = g_m^{\ell,h} A_{i,s_m}^{\ell,h}, \qquad \widetilde{A}_{i,j}^{\ell,h} = A_{i,j}^{\ell,h} + \Delta_i^{\ell,h} r_{i,j}^{\ell,h}, \quad j \in \mathcal{V}_i^-. \tag{20}$$

Since all released sink attention is reassigned to visual non-sink tokens, the total attention mass is preserved. The redistributed attention $\widetilde{A}^{\ell,h}$ is then used in standard multi-head attention. Alternative redistribution strategies are compared in Sec. 6.2.

5.3 TRAINING

We freeze the pretrained LVLM and optimize only the simple gating modules using the standard next-token prediction objective:

$$\mathcal{L}_{\mathrm{NTP}} = -\sum_{t} \log p_\theta \left(y_t \mid X, y_{<t} \right), \tag{21}$$

where θ denotes the gating parameters. To prevent excessive sink suppression from disrupting the anchoring structure (Sec. 4.3), we introduce a soft retention constraint. For each layer, head, and query with non-zero sink attention, we define

$$R_i^{\ell,h} = \frac{\sum_{m=1}^{K} g_m^{\ell,h} \ \mathrm{sg}\Big(A_{i,s_m}^{\ell,h}\Big)}{\sum_{m=1}^{K} \mathrm{sg}\Big(A_{i,s_m}^{\ell,h}\Big) + \epsilon}, \tag{22}$$

where $\mathrm{sg}(\cdot)$ denotes stop-gradient. We penalize only retention ratios below a threshold τ:

$$\mathcal{L}_{\mathrm{ret}} = \frac{\sum_{\ell,h} \sum_{i \in \mathcal{Q}_{\mathcal{S}}^{\ell,h}} \Big[\max(0, \tau - R_i^{\ell,h})\Big]^2}{\sum_{\ell,h} |\mathcal{Q}_{\mathcal{S}}^{\ell,h}|}, \tag{23}$$

where $\mathcal{Q}_{\mathcal{S}}^{\ell,h}$ denotes valid queries with non-zero sink attention. The overall objective is

$$\mathcal{L} = \mathcal{L}_{\mathrm{NTP}} + \lambda_{\mathrm{ret}} \mathcal{L}_{\mathrm{ret}}. \tag{24}$$

We initialize the gates near the identity mapping ($g_m^{\ell,h} \approx 0.95$), allowing training to start from the pretrained behavior while avoiding excessive sink suppression.

6 EXPERIMENTS

6.1 EXPERIMENTAL SETUP

Models and Baselines. We evaluate Qwen3-VL-8B, InternVL3.5-8B, and LLaVA-1.5-7B. Baselines include the vanilla models, hallucination mitigation methods that suppress or remove AS (OPERA (Huang et al., 2024) and TAME (Tang et al., 2025a)), and AS regulation methods (VAR (Kang et al., 2025), DIYSink (Luo et al., 2025), and LSG (Choi et al., 2026)).

Evaluation Benchmarks. We evaluate three capability dimensions: **general visual understanding** (GQA, TextVQA, MMMU, MME, and MM-Vet), **hallucination** (POPE and CHAIR), and **visual reasoning** (ScienceQA and MathVista). We follow the standard metrics of each benchmark; MME is reported as MME_P and MME_C, and lower CHAIR-S/I indicates fewer hallucinations.

Implementation Details. We set `max_new_tokens`=128 and use consistent inference settings within each task category. We freeze the pretrained backbones and train only the gating modules on 10K stratified Cambrian-7M samples for 10 epochs with the standard next-token prediction objective, initializing $g \approx 0.95$ and setting $\tau = 0.2$ and $\lambda_{\mathrm{ret}} = 10$. The best checkpoint is selected on the validation set. All baselines use their official implementations and default configurations. Detailed experimental settings and hyperparameter sensitivity analyses are provided in the Appendix B.

6.2 MAIN RESULTS

Consistent gains across models and capabilities. As shown in Table 4, G-SAR consistently improves over Greedy across all 39 model–metric combinations and achieves the best or tied-best result on 29 of them. The improvements span all three capability dimensions and remain consistent across Qwen3-VL-8B, InternVL3.5-8B, and LLaVA-1.5-7B. In particular, G-SAR improves MathVista by $+3.5$, $+3.1$, and $+4.5$ points, respectively, while also yielding gains of up to $+26.7$ on MME-P and $+18.1$ on MME-C. This cross-model and cross-capability consistency suggests that sink regulation enhances the utilization of visual evidence without sacrificing the model's original capabilities.

Regulation alleviates the hallucination–reasoning trade-off. Table 4 shows that existing hallucination mitigation methods can achieve strong reductions in CHAIR scores, but often at the cost

Table 4: **Performance comparison across general visual, hallucination, and visual reasoning benchmarks.** Lower is better for CHAIR-S/I and higher otherwise. Bold denotes the best result within each model block; superscripts indicate changes from the corresponding Greedy baseline.

Methods	General Visual Tasks						Hallucination Tasks					Visual Reasoning Tasks	
	GQA	TextVQA	MMMU	MME_P	MME_C	MM-Vet	POPE-R	POPE-P	POPE-A	CHAIR-S ↓	CHAIR-I ↓	SciQA	MathVista
Qwen3-VL-8B													
Greedy	72.4	82.7	29.0	1716.7	613.9	66.9	92.7	90.1	90.5	24.3	6.8	68.0	37.4
OPERA	72.3$_{-0.1}$	83.0$_{+0.3}$	29.1$_{+0.1}$	1723.7$_{+7.0}$	620.0$_{+6.1}$	65.1$_{-1.8}$	92.7$_{0.0}$	90.4$_{+0.3}$	90.5$_{0.0}$	22.5$_{+1.8}$	6.2$_{+0.6}$	63.0$_{-5.0}$	36.0$_{-1.4}$
TAME	71.7$_{-0.7}$	82.9$_{+0.2}$	28.2$_{-0.8}$	1686.7$_{-30.0}$	563.9$_{-50.0}$	66.3$_{-0.6}$	92.9$_{+0.2}$	90.6$_{+0.5}$	90.7$_{+0.2}$	20.8$_{+3.5}$	**5.3**$_{+1.5}$	65.9$_{-2.1}$	33.4$_{-4.0}$
VAR	72.3$_{-0.1}$	82.7$_{0.0}$	29.2$_{+0.2}$	**1738.0**$_{+21.3}$	**622.9**$_{+9.0}$	68.2$_{+1.3}$	92.8$_{+0.1}$	90.1$_{0.0}$	90.3$_{-0.2}$	24.7$_{-0.4}$	7.0$_{-0.2}$	68.5$_{+0.5}$	38.2$_{+0.8}$
DIYSink	73.0$_{+0.6}$	82.0$_{-0.7}$	29.2$_{+0.2}$	1724.7$_{+8.0}$	615.9$_{+2.0}$	68.0$_{+1.1}$	92.7$_{0.0}$	90.4$_{+0.3}$	90.7$_{+0.2}$	23.8$_{+0.5}$	6.5$_{+0.3}$	69.0$_{+1.0}$	39.9$_{+2.5}$
LSG	73.9$_{+1.5}$	83.1$_{+0.4}$	29.6$_{+0.6}$	1729.3$_{+12.6}$	614.4$_{+0.5}$	**69.0**$_{+2.1}$	92.5$_{-0.2}$	90.1$_{0.0}$	90.6$_{+0.1}$	24.0$_{+0.3}$	6.6$_{+0.2}$	68.1$_{+0.1}$	40.0$_{+2.6}$
G-SAR (Ours)	**74.2**$_{+1.8}$	**83.7**$_{+1.0}$	**29.9**$_{+0.9}$	1735.0$_{+18.3}$	619.9$_{+6.0}$	68.7$_{+1.8}$	**93.2**$_{+0.5}$	**90.9**$_{+0.8}$	**90.9**$_{+0.4}$	**20.7**$_{+3.6}$	5.8$_{+1.0}$	**70.8**$_{+2.8}$	**40.9**$_{+3.5}$
InternVL3.5-8B													
Greedy	71.0	77.1	53.0	1636.1	622.2	62.9	89.3	87.6	87.6	25.1	7.7	95.2	40.0
OPERA	70.4$_{-0.6}$	77.4$_{+0.3}$	52.8$_{-0.2}$	1642.6$_{+6.5}$	618.7$_{-3.5}$	60.7$_{-2.2}$	89.4$_{+0.1}$	**88.4**$_{+0.8}$	88.1$_{+0.5}$	**20.5**$_{+4.6}$	**5.9**$_{+1.8}$	91.8$_{-3.4}$	38.1$_{-1.9}$
TAME	70.0$_{-1.0}$	76.3$_{-0.8}$	52.2$_{-0.8}$	1594.8$_{-41.3}$	557.6$_{-64.6}$	61.9$_{-1.0}$	89.6$_{+0.3}$	88.1$_{+0.5}$	88.3$_{+0.7}$	21.7$_{+3.4}$	6.5$_{+1.2}$	90.7$_{-4.5}$	35.2$_{-4.8}$
VAR	71.7$_{+0.7}$	77.5$_{+0.4}$	53.4$_{+0.4}$	1653.6$_{+17.5}$	631.6$_{+9.4}$	64.2$_{+1.3}$	89.1$_{-0.2}$	87.9$_{+0.3}$	87.7$_{+0.1}$	23.6$_{+1.5}$	7.3$_{+0.4}$	94.8$_{-0.4}$	41.3$_{+1.3}$
DIYSink	71.4$_{+0.4}$	76.8$_{-0.3}$	53.3$_{+0.3}$	1649.4$_{+13.3}$	625.6$_{+3.4}$	64.0$_{+1.1}$	89.3$_{0.0}$	87.7$_{+0.1}$	87.8$_{+0.2}$	24.1$_{+1.0}$	7.2$_{+0.5}$	95.8$_{+0.6}$	41.9$_{+1.9}$
LSG	72.0$_{+1.0}$	77.4$_{+0.3}$	53.2$_{+0.2}$	1644.6$_{+8.5}$	623.5$_{+1.3}$	64.1$_{+1.2}$	89.4$_{+0.1}$	87.5$_{-0.1}$	87.7$_{+0.1}$	24.9$_{+0.2}$	7.7$_{0.0}$	95.7$_{+0.5}$	41.2$_{+1.2}$
G-SAR (Ours)	**72.9**$_{+1.9}$	**78.1**$_{+1.0}$	**54.1**$_{+1.1}$	**1662.8**$_{+26.7}$	**638.4**$_{+16.2}$	**65.2**$_{+2.3}$	**89.8**$_{+0.5}$	88.3$_{+0.7}$	**88.4**$_{+0.8}$	20.8$_{+4.3}$	6.3$_{+1.4}$	**97.0**$_{+1.8}$	**43.1**$_{+3.1}$
LLaVA-1.5-7B													
Greedy	61.9	58.2	27.3	1466.9	282.9	31.0	88.3	87.0	84.7	51.0	14.3	64.8	24.4
OPERA	61.2$_{-0.7}$	58.3$_{+0.1}$	27.3$_{0.0}$	1464.5$_{-2.4}$	279.1$_{-3.8}$	28.4$_{-2.6}$	88.5$_{+0.2}$	**88.2**$_{+1.2}$	85.9$_{+1.2}$	**44.1**$_{+6.9}$	12.8$_{+1.5}$	62.0$_{-2.8}$	22.4$_{-2.0}$
TAME	60.9$_{-1.0}$	56.9$_{-1.3}$	26.1$_{-1.2}$	1423.8$_{-43.1}$	204.6$_{-78.3}$	30.1$_{-0.9}$	**88.6**$_{+0.3}$	87.8$_{+0.8}$	85.8$_{+1.1}$	48.9$_{+2.1}$	**12.1**$_{+2.2}$	58.9$_{-5.9}$	18.6$_{-5.8}$
VAR	**63.5**$_{+1.6}$	58.6$_{+0.4}$	27.6$_{+0.3}$	1478.3$_{+11.4}$	293.5$_{+10.6}$	31.8$_{+0.8}$	88.0$_{-0.3}$	87.2$_{+0.2}$	85.0$_{+0.3}$	48.1$_{+2.9}$	13.7$_{+0.6}$	64.1$_{-0.7}$	25.6$_{+1.2}$
DIYSink	62.1$_{+0.2}$	58.3$_{+0.1}$	27.4$_{+0.1}$	1480.5$_{+13.6}$	285.4$_{+2.5}$	32.0$_{+1.0}$	88.2$_{-0.1}$	87.0$_{0.0}$	84.7$_{0.0}$	50.3$_{+0.7}$	13.9$_{+0.4}$	65.2$_{+0.4}$	25.1$_{+0.7}$
LSG	62.8$_{+0.9}$	58.2$_{0.0}$	27.4$_{+0.1}$	1472.1$_{+5.2}$	283.6$_{+0.7}$	31.1$_{+0.1}$	88.3$_{0.0}$	87.0$_{0.0}$	84.7$_{0.0}$	51.2$_{-0.2}$	14.4$_{-0.1}$	65.0$_{+0.2}$	24.9$_{+0.5}$
G-SAR (Ours)	**63.5**$_{+1.6}$	**59.2**$_{+1.0}$	**27.9**$_{+0.6}$	**1492.6**$_{+25.7}$	**301.0**$_{+18.1}$	**32.8**$_{+1.8}$	**88.6**$_{+0.3}$	87.8$_{+0.8}$	**86.0**$_{+1.3}$	46.6$_{+4.4}$	12.8$_{+1.5}$	**65.3**$_{+0.5}$	**28.9**$_{+4.5}$

Table 5: **Ablation study of attention redistribution targets on Qwen3-VL-8B.** Values indicate improvements over the Greedy. Bold and underlined values denote the best and second results.

Redistribution	ΔGQA	ΔMME	ΔPOPE-A	ΔCHAIR-S	ΔCHAIR-I	ΔMathVista
All Tokens	+0.4	+6.0	+0.0	+0.8	+0.2	+0.6
All Non-Sink Tokens	+0.8	+11.0	+0.1	+1.6	+0.4	+1.3
All Visual Tokens	+1.2	+17.0	<u>+0.2</u>	<u>+2.1</u>	<u>+0.6</u>	+2.1
Query-Guided Visual Tokens	**+2.0**	**+26.0**	+0.1	-2.4	-0.5	<u>+3.5</u>
Visual Non-Sink (G-SAR)	<u>+1.8</u>	<u>+24.3</u>	**+0.4**	**+3.6**	**+1.0**	**+3.5**

of reasoning degradation. For example, across ScienceQA and MathVista, OPERA and TAME decrease reasoning performance by an average of 2.75 and 4.52 points relative to Greedy, respectively. In contrast, G-SAR improves reasoning by an average of 2.70 points while reducing CHAIR-S and CHAIR-I by 3.6–4.4 and 1.0–1.5 points, respectively, across the three models. These results suggest that the key issue is not the presence of AS themselves, but their excessive concentration: adaptively redistributing sink attention can suppress hallucination while improving visual reasoning capability.

Effect of Redistribution Strategies. Table 5 shows that the redistribution target is critical. Redirecting released attention to visual tokens generally improves performance, while Query-Guided Visual Tokens further boosts understanding and reasoning but degrades hallucination metrics, suggesting that query relevance alone may over-concentrate attention on limited visual evidence. In contrast, G-SAR redistributes attention to visual non-sink tokens, achieving the strongest hallucination improvements (+0.4 POPE-A, +3.6 CHAIR-S, and +1.0 CHAIR-I) while retaining near-best gains in understanding and reasoning (+1.8 GQA, +24.3 MME, and +3.5 MathVista). These results suggest that effective redistribution should enhance visual evidence without reinforcing already dominant tokens. Similar trends on LLaVA-1.5-7B and InternVL3.5-8B are provided in Appendix B.5.

7 Conclusion

In this work, we uncover an underexplored anchoring effect of AS in LVLMs. We show that AS persist across layers and heads and structurally support attention organization and representation geometry, while excessive suppression can reduce hallucination at the cost of visual reasoning. Motivated by this dual role, we propose G-SAR, a simple yet effective adaptive mechanism that regulates sink attention by redistributing excess mass to visual non-sink tokens. Across three LVLMs and diverse benchmarks, G-SAR reduces hallucination while improving visual understanding and reasoning.

AI USE STATEMENT

Generative AI tools were used only for language polishing and writing suggestions during manuscript preparation. All AI-assisted text was manually reviewed, revised, and proofread by the authors. AI tools were not used for problem formulation, experimental design, result interpretation, or drawing conclusions. The authors take full responsibility for the final manuscript.

ETHICAL CONSIDERATIONS

This work studies the internal attention behavior of large vision-language models and proposes a simple mechanism for regulating attention sinks. The study focuses on model analysis, hallucination mitigation, and balancing visual understanding and reasoning, without introducing new high-risk capabilities or collecting sensitive data. The proposed method is intended for authorized model analysis, evaluation, and reliability research, and model outputs should be handled with appropriate caution.

REPRODUCIBILITY STATEMENT

Section 3 defines attention sinks, while Section 4 introduces the metrics used to characterize their anchoring effect. Section 5 details the architecture, attention redistribution mechanism, and training objective of G-SAR. Section 6 and Appendix B detail the experimental setup, including baselines, benchmarks, training data, hyperparameters, and implementation settings.

REFERENCES

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikoł aj Bińkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. Flamingo: a visual language model for few-shot learning. In *NIPS*, pp. 23716–23736, 2022.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, et al. Qwen3-vl technical report, 2025. URL https://arxiv.org/abs/2511.21631.

Federico Barbero, Alvaro Arroyo, Xiangming Gu, Christos Perivolaropoulos, Petar Veličković, Razvan Pascanu, and Michael M. Bronstein. Why do LLMs attend to the first token? In *COLM*, 2025.

Kai Chen, Yanze Li, Wenhua Zhang, Yanxin Liu, Pengxiang Li, Ruiyuan Gao, Lanqing Hong, Meng Tian, Xinhai Zhao, Zhenguo Li, Dit-Yan Yeung, Huchuan Lu, and Xu Jia. Automated evaluation of large vision-language models on self-driving corner cases. In *WACV*, pp. 7817–7826, 2025a.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*, pp. 19–35, 2025b.

Xuweiyi Chen, Ziqiao Ma, Xuejun Zhang, Sihan Xu, Shengyi Qian, Jianing Yang, David Fouhey, and Joyce Chai. Multi-object hallucination in vision language models. In *NIPS*, 2024.

Jiho Choi, Jaemin Kim, Sanghwan Kim, Seunghoon Hong, and Jin-Hwi Park. When sinks help or hurt: Unified framework for attention sink in large vision-language models. In Paolo Favaro, Zuzana Kukelova, Atsuto Maki, Anna Rohrbach, Konrad Schindler, and Federico Tombari (eds.), *ECCV*, pp. 148–166, 2026.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale Fung, and Steven Hoi. InstructBLIP: Towards general-purpose vision-language models with instruction tuning. In *NIPS*, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. In *ICLR*, pp. 2632–2652, 2024.

Enrique Queipo de Llano, Álvaro Arroyo, Federico Barbero, Xiaowen Dong, Michael Bronstein, Yann LeCun, and Ravid Shwartz-Ziv. Attention sinks and compression valleys in llms are two sides of the same coin, 2026. URL https://arxiv.org/abs/2510.06477.

Yuchen Duan, Zhe Chen, Yusong Hu, Weiyun Wang, Shenglong Ye, Botian Shi, Lewei Lu, Qibin Hou, Tong Lu, Hongsheng Li, et al. Docopilot: Improving multimodal models for document-level understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 4026–4037, 2025.

Suyu Ge, Yunan Zhang, Liyuan Liu, Minjia Zhang, Jiawei Han, and Jianfeng Gao. Model tells you what to discard: Adaptive KV cache compression for LLMs. In *ICLR*, 2024.

Xiangming Gu, Tianyu Pang, Chao Du, Qian Liu, Fengzhuo Zhang, Cunxiao Du, Ye Wang, and Min Lin. When attention sink emerges in language models: An empirical view. In *ICLR*, 2025.

Chi Han, Qifan Wang, Hao Peng, Wenhan Xiong, Yu Chen, Heng Ji, and Sinong Wang. Lm-infinite: Zero-shot extreme length generalization for large language models. In *NAACL*, pp. 3991–4008, 2024.

Qidong Huang, Xiaoyi Dong, Pan Zhang, Bin Wang, Conghui He, Jiaqi Wang, Dahua Lin, Weiming Zhang, and Nenghai Yu. Opera: Alleviating hallucination in multi-modal large language models via over-trust penalty and retrospection-allocation. In *CVPR*, pp. 13418–13427, 2024.

Linzhao Jia, Han Sun, Jialong Jiang, and Xiaozhe Yang. High-quality classroom dialogue automatic analysis system. *Applied Sciences*, 15(3), 2025. ISSN 2076-3417. doi: 10.3390/app15031613. URL https://www.mdpi.com/2076-3417/15/3/1613.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. In *ICLR*, 2025.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models, 2023. URL https://arxiv.org/abs/2301.12597.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *NIPS*, pp. 34892–34916, 2023.

Jiayun Luo, Wan-Cyuan Fan, Lyuyang Wang, Xiangteng He, Tanzila Rahman, Purang Abolmaesumi, and Leonid Sigal. To sink or not to sink: Visual information pathways in large vision-language models, 2025. URL https://arxiv.org/abs/2510.08510.

Xinyin Ma, Gongfan Fang, and Xinchao Wang. Llm-pruner: On the structural pruning of large language models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *NIPS*, pp. 21702–21720, 2023.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, Qixiang Ye, and Furu Wei. Grounding multimodal large language models to the world. In *ICLR*, 2024.

Valeria Ruscio, Umberto Nanni, and Fabrizio Silvestri. What are you sinking? a geometric approach on attention sink. In *NIPS*, 2025.

Pedro Sandoval-Segura, Xijun Wang, Ashwinee Panda, Micah Goldblum, Ronen Basri, Tom Goldstein, and David Jacobs. Identifying and evaluating inactive heads in pretrained LLMs. In *ICLR*, 2026.

Yongliang Shen, Kaitao Song, Xu Tan, Dongsheng Li, Weiming Lu, and Yueting Zhuang. Hugginggpt: Solving ai tasks with chatgpt and its friends in hugging face, 2023. URL https://arxiv.org/abs/2303.17580.

Chonghao Sima, Katrin Renz, Kashyap Chitta, Li Chen, Hanxue Zhang, Chengen Xie, Jens Beißwenger, Ping Luo, Andreas Geiger, and Hongyang Li. Drivelm: Driving with graph visual question answering. In *ECCV*, pp. 256–274, 2025.

Han Sun, Qin Li, Peixin Wang, and Min Zhang. Mitigating object hallucinations in lvlms via attention imbalance rectification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Findings*, pp. 8930–8940, June 2026a.

Shangwen Sun, Alfredo Canziani, Yann LeCun, and Jiachen Zhu. The spike, the sparse and the sink: Anatomy of massive activations and attention sinks, 2026b. URL https://arxiv.org/abs/2603.05498.

Rinyoichi Takezoe, Yaqian Li, Zi-Hao Bo, Anzhou Hou, Mo Guang, and Kaiwen Long. Learnpruner: Rethinking attention-based token pruning in vision language models. In *ICLR*, 2026.

Feilong Tang, Zile Huang, Chengzhi Liu, Qiang Sun, Harry Yang, and Ser-Nam Lim. Intervening anchor token: Decoding strategy in alleviating hallucinations for MLLMs. In *ICLR*, 2025a.

Feilong Tang, Chengzhi Liu, Zhongxing Xu, Ming Hu, Zile Huang, Haochen Xue, Ziyang Chen, Zelin Peng, Zhiwei Yang, Sijin Zhou, Wenxue Li, Yulong Li, Wenxuan Song, Shiyan Su, Wei Feng, Jionglong Su, Minquan Lin, Yifan Peng, Xuelian Cheng, Imran Razzak, and Zongyuan Ge. Seeing far and clearly: Mitigating hallucinations in mllms with attention causal decoding. In *CVPR*, pp. 26147–26159, 2025b.

Jue Wang. Hallucination reduction and optimization for large language model-based autonomous driving. *Symmetry*, 16(9), 2024.

Lean Wang, Lei Li, Damai Dai, Deli Chen, Hao Zhou, Fandong Meng, Jie Zhou, and Xu Sun. Label words are anchors: An information flow perspective for understanding in-context learning. In *EMNLP*, pp. 9840–9855, 2023.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, et al. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency, 2025a. URL https://arxiv.org/abs/2508.18265.

Yining Wang, Mi Zhang, Junjie Sun, Chenyue Wang, Min Yang, Hui Xue, Jialing Tao, Ranjie Duan, and Jiexi Liu. Mirage in the eyes: Hallucination attack on multi-modal large language models with only attention sink. In *USENIX Security Symposium*, 2025b.

Haoyi Wu and Kewei Tu. Layer-condensed KV cache for efficient inference of large language models. In *ACL*, pp. 11175–11188, 2024.

Xixi Wu, Yanchao Tan, Nan Hou, Ruiyang Zhang, and Hong Cheng. MoLoRAG: Bootstrapping document understanding via multi-modal logic-aware retrieval. In *EMNLP*, pp. 14024–14045, 2025.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. In *ICLR*, volume 2024, pp. 21875–21895, 2024.

Shuai Yang, Yuying Ge, Yang Li, Yukang Chen, Yixiao Ge, Ying Shan, and Yingcong Chen. Seedstory: Multimodal long story generation with large language model. In *ICCVW*, pp. 1871–1881, 2025.

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. Atp-llava: Adaptive token pruning for large vision language models. In *CVPR*, pp. 24972–24982, 2025.

Stephen Zhang, Mustafa Khan, and Vardan Papyan. Attention sinks: A 'catch, tag, release'mechanism for embeddings. In *NIPS*, pp. 83140–83181, 2025a.

Xiaofeng Zhang, Yihao Quan, Chen Shen, Chaochen Gu, Xiaosong Yuan, Shaotian Yan, Jiawei Cao, Hao Cheng, Kaijie Wu, and Jieping Ye. Shallow focus, deep fixes: Enhancing shallow layers vision attention sinks to alleviate hallucination in LVLMs. In *EMNLP*, pp. 3512–3534, 2025b.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *ICML*, 2025c.

Zhenyu Zhang, Shiwei Liu, Runjin Chen, Bhavya Kailkhura, Beidi Chen, and Zhangyang Wang. Q-hitter: A better token oracle for efficient llm inference via sparse-quantized kv cache. In *MLSys*, volume 6, pp. 381–394, 2024.

A Additional Analysis

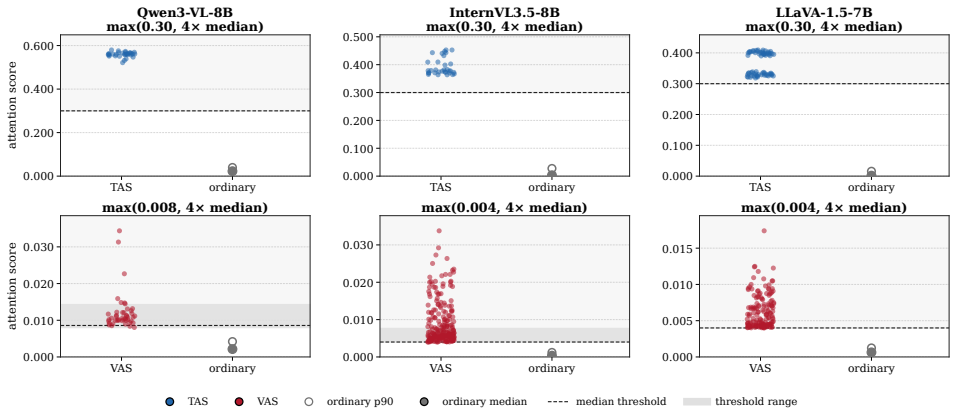

Figure 2: **Attention distributions under the AS identification criterion.** The top and bottom rows show TAS and VAS across three LVLMs. Colored points denote detected sink-token attention scores; gray markers show the median and 90th percentile of non-sink tokens, while dashed lines and shaded bands indicate the median and range of adaptive thresholds. The two populations exhibit a clear empirical separation under the adopted criterion.

A.1 Attention Sink Identification

We first perform an intervention-free prefill to measure the incoming attention received by each token. Since textual and visual tokens operate at markedly different attention scales, we identify attention sinks using modality-specific adaptive thresholds.

Sink threshold. For modality $m \in \{\text{txt}, \text{vis}\}$, we define

$$\tau_m = \max(\tau_{0,m},\, 4M_m),$$

where M_m denotes the median non-zero incoming attention within the corresponding modality and $\tau_{0,m}$ is a minimum threshold floor. The median-dependent term adapts the criterion to the instance-specific background attention level, while the floor prevents unusually small medians from producing overly permissive thresholds.

A token is identified as an attention sink only if at least a fraction $\gamma = 0.5$ of its causally valid queries assign attention above the corresponding threshold. This joint magnitude-and-persistence criterion filters out isolated high-attention events and retains tokens that exhibit consistently elevated incoming attention.

For textual tokens, we use $\tau_{\text{txt}} = \max(0.30,\, 4M_{\text{txt}})$ for all models. For visual tokens, we use $\tau_{\text{vis}} = \max(0.008,\, 4M_{\text{vis}})$ for Qwen3-VL-8B and $\tau_{\text{vis}} = \max(0.004,\, 4M_{\text{vis}})$ for InternVL3.5-8B and LLaVA-1.5-7B.

Figure 2 further characterizes the attention distributions induced by this identification rule. Here, ordinary tokens refer to all tokens that do not satisfy the above AS identification criterion (non-sink tokens). Across all three LVLMs, detected sinks occupy the high-attention tail, while ordinary tokens remain concentrated at substantially lower attention levels. In particular, the 90th percentile of ordinary-token attention remains below the corresponding sink threshold in all settings. This analysis characterizes the empirical separation between detected sink and non-sink tokens under the adopted criterion.

TAS and VAS exhibit distinct positional patterns. Figure 3 reveals a clear contrast between textual and visual attention sinks across the 2,374 samples in the MME dataset. TAS are highly stable and concentrated at a small number of fixed prompt positions. For Qwen3-VL-8B and InternVL3.5-8B, TAS almost exclusively occur at the image-start token, while LLaVA-1.5-7B consistently exhibits TAS at two fixed textual positions. This strong positional regularity suggests that TAS are largely

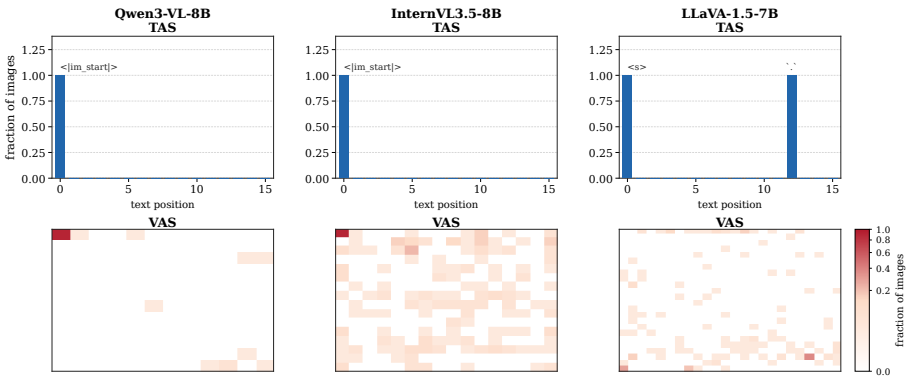

Figure 3: **Positional patterns of textual and visual attention sinks across LVLMs.** The top row shows the occurrence frequency of TAS at each text position, while the bottom row shows the spatial distribution of VAS over image patches. ***Key findings:*** TAS are concentrated at fixed prompt positions, whereas VAS exhibit distinct model-specific spatial patterns.

governed by sequence structure—such as causal masking and the placement of special prompt tokens—rather than by image-specific semantic content, thereby serving as stable anchors in the text stream. This observation is consistent with prior findings Barbero et al. (2025); Gu et al. (2025); Ruscio et al. (2025).

In contrast, VAS exhibit clear model-dependent spatial patterns. Qwen3-VL-8B shows a sharply localized distribution, with sinks concentrated on the first visual token and only limited spillover to a few other patch positions. InternVL3.5-8B displays the most diffuse pattern, with sinks spread across both border and interior patches. LLaVA-1.5-7B exhibits a different inductive bias, with sinks occurring predominantly on border patches, especially near the bottom edge. Overall, whereas TAS are template-like and nearly fixed across samples, VAS are more architecture-dependent and reflect how different LVLMs organize visual tokens.

B Additional Experiments

B.1 Baseline Details

Vanilla. We use the original LVLMs without any attention intervention or modified decoding strategy, serving as the reference for evaluating the effect of attention-sink regulation.

OPERA (Huang et al., 2024). OPERA is a training-free decoding method for hallucination mitigation. It attributes hallucination partly to excessive reliance on a few summary tokens and introduces an *over-trust penalty* during beam-search decoding to discourage such concentrated attention. A retrospection-allocation mechanism further rolls back and reallocates token selection when excessive aggregation is detected.

TAME (Tang et al., 2025a). TAME targets excessive propagation through anchor tokens, which can cause the model to over-rely on textual summaries while neglecting visual evidence. Instead of directly masking these tokens, it dynamically intervenes in the eigenspectrum of the query–key interaction to reduce excessive attention localization, providing a plug-and-play hallucination mitigation strategy without additional training.

VAR (Kang et al., 2025). Visual Attention Redistribution (VAR) identifies visual attention sinks that consistently receive high attention despite limited relevance to the current query. It treats the attention assigned to such sinks as surplus and redistributes it within image-centric attention heads toward informative visual tokens, improving visual utilization without additional training.

DIYSink (Luo et al., 2025). DIYSink builds on the observation that visual sinks can originate from the vision encoder and may encode useful global information. It explicitly distinguishes sink and

non-sink visual pathways and adaptively controls their relative contributions, enabling the model to preserve global information carried by sink tokens while retaining fine-grained visual evidence.

LSG (Choi et al., 2026). Layer-wise Sink Gating (LSG) distinguishes vision-encoder-emerged sinks from sinks arising within the language model and regulates their contributions in a layer-dependent manner. A simple learned gate dynamically scales sink and non-sink visual attention while keeping the pretrained LVLM backbone frozen, balancing global visual priors with local visual evidence.

Comparison with prior sink-regulation methods. G-SAR differs from prior approaches by coupling adaptive sink regulation with explicit attention redistribution. VAR redistributes a fixed fraction of visual-sink attention, whereas G-SAR learns input-adaptive, sink- and head-specific gates to determine how much attention to retain. DIYSink regulates sink information through separate vision–language pathways, while G-SAR directly operates on LLM attention without modifying the visual projector. LSG performs layer-wise group-level key scaling, whereas G-SAR further reassigns the released attention mass to visual non-sink tokens using the pretrained query–visual attention pattern as a routing prior. This mass-conserving redistribution distinguishes G-SAR from methods that primarily suppress or rescale sink attention: it jointly controls ***how much*** attention remains on sinks and ***where*** the released attention is routed. Motivated by the anchoring role of AS, G-SAR therefore regulates rather than indiscriminately removes sink attention, alleviating excessive concentration while retaining their structural contribution.

B.2 Evaluation Metrics

General Visual Understanding. We evaluate general vision–language understanding on GQA, TextVQA, MMMU, MME, and MM-Vet, which collectively cover visual question answering, text-centric perception, multi-disciplinary reasoning, and comprehensive multimodal understanding. For GQA, we report the standard answer accuracy over compositional visual questions. TextVQA is evaluated using its official VQA accuracy metric, which measures agreement between model predictions and human-annotated answers for questions requiring scene-text understanding. For MMMU, we report the standard accuracy over questions spanning multiple academic disciplines and heterogeneous visual inputs.

MME evaluates multimodal capabilities along two dimensions: **perception** and **cognition**. Following the official evaluation protocol, we report the accumulated scores over the corresponding subtasks separately as **MME_P** and **MME_C**, where higher scores indicate stronger perceptual and cognitive capabilities, respectively. MM-Vet assesses integrated vision–language capabilities, including recognition, spatial understanding, knowledge, reasoning, and generation, using its official GPT-based evaluator. We report the resulting MM-Vet score, with higher values indicating better overall multimodal performance.

Hallucination. We evaluate object hallucination using POPE and CHAIR. POPE formulates hallucination evaluation as a binary visual question answering task under three sampling strategies: *Random*, *Popular*, and *Adversarial*. We report the accuracy on each split, with higher scores indicating a lower degree of visual hallucination.

CHAIR evaluates hallucinated object mentions in free-form image captions. We prompt each LVLM with "*Please describe this image in detail.*" and report sentence-level and instance-level hallucination rates:

$$\mathrm{CHAIR_S} = \frac{N_{\mathrm{cap}}^{\mathrm{hall}}}{N_{\mathrm{cap}}},$$

$$\mathrm{CHAIR_I} = \frac{N_{\mathrm{obj}}^{\mathrm{hall}}}{N_{\mathrm{obj}}},$$

where $N_{\mathrm{cap}}^{\mathrm{hall}}$ denotes the number of captions containing at least one hallucinated object, N_{cap} the total number of generated captions, $N_{\mathrm{obj}}^{\mathrm{hall}}$ the number of hallucinated object mentions, and N_{obj} the total number of object mentions. Lower $\mathrm{CHAIR_S}$ and $\mathrm{CHAIR_I}$ indicate fewer hallucinations.

Table 6: **Detailed training configuration of G-SAR.**

Configuration	Value
G-SAR configuration	
Gating MLP hidden dimension	64
Routing configuration	`split`
Sink-frequency threshold γ	0.5
Retention threshold τ	0.2
Retention weight λ_{ret}	10
Optimization configuration	
Training samples	10K
Training epochs	10
Learning rate	1×10^{-3}
Learning-rate schedule	Cosine decay
Warmup ratio	0.03
Per-device batch size	2
Number of GPUs	2
Gradient accumulation steps	4
Effective batch size	16
Maximum sequence length	2,048
Maximum image pixels	902,816
Precision	BF16
Gradient checkpointing	Enabled
Hardware	2× NVIDIA A40 (50 GB)

Visual Reasoning. We evaluate visual reasoning on ScienceQA and MathVista, which cover complementary forms of multimodal reasoning. ScienceQA assesses scientific reasoning over multimodal questions involving visual content, domain knowledge, and relational inference, while MathVista focuses on mathematical reasoning grounded in visual inputs, including numerical calculation, quantitative comparison, and multi-step problem solving. Following the standard evaluation protocols of both benchmarks, we report answer accuracy, with higher values indicating stronger visual reasoning capability.

B.3 Training Details

We freeze all parameters of the pretrained LVLM backbone and optimize only the simple G-SAR gating modules. Following the training data construction in prior sink-gating work Choi et al. (2026), we use 10K samples from Cambrian-7M, stratified equally across six task categories: OCR, general VQA, counting, code, math, and science, with approximately 1.67K samples per category. This balanced subset provides diverse visual and reasoning supervision while avoiding task-specific downstream training. We train G-SAR for 10 epochs, with a separate gating MLP instantiated for each Transformer layer and attention head. The detailed training configuration is summarized in Table 6.

The sink-frequency threshold is fixed at $\gamma = 0.5$. For the retention regularization, we set $\tau_{\text{ret}} = 0.2$ and $\lambda_{\text{ret}} = 10$, where τ_{ret} specifies the soft lower bound on the sink retention ratio and λ_{ret} controls the strength of the corresponding penalty. These settings prevent excessive suppression of sink attention while retaining sufficient flexibility for adaptive redistribution. All experiments are conducted on two NVIDIA A40 GPUs, each with 50 GB of memory.

B.4 Sensitivity to Joint Retention Strength

We examine the sensitivity of G-SAR to the overall strength of the retention constraint by jointly varying the retention threshold τ_{ret} and the regularization weight λ_{ret} along four configurations: $(\tau_{\text{ret}}, \lambda_{\text{ret}}) \in \{(0.1, 5), (0.2, 10), (0.4, 15), (0.6, 20)\}$, with $(0.2, 10)$ used as the default setting. Since the two hyperparameters are varied jointly, this experiment evaluates sensitivity to the overall retention strength rather than isolating their individual effects.

Table 7: **Sensitivity of G-SAR to joint retention strength on InternVL3.5-8B.** We jointly vary τ_{ret} and λ_{ret} from weak to strong retention constraints. Values are changes relative to Greedy; CHAIR-S/I signs are reversed so that higher is better. Bold and underline denote the best and second-best results.

τ_{ret}	λ_{ret}	ΔGQA	ΔMME	ΔPOPE-A	ΔCHAIR-S	ΔCHAIR-I	ΔMathVista
0.1	5	<u>+1.8</u>	+12.4	+0.1	+3.9	+1.0	+1.8
0.2	10	**+1.9**	**+42.9**	**+0.8**	**+4.3**	**+1.4**	+3.1
0.4	15	+0.3	<u>+33.9</u>	<u>+0.5</u>	<u>+4.0</u>	<u>+1.1</u>	<u>+3.3</u>
0.6	20	-0.1	+19.4	+0.0	+1.1	+0.6	**+5.0**

Table 8: **Additional ablation of attention redistribution targets on InternVL3.5-8B.** Values report performance changes relative to the Greedy baseline. Positive values consistently denote improvements; for CHAIR-S and CHAIR-I, the sign is reversed such that larger values indicate greater reductions in hallucination. Bold and underline indicate the best and second-best results, respectively.

Redistribution	ΔGQA	ΔMME	ΔPOPE-A	ΔCHAIR-S	ΔCHAIR-I	ΔMathVista
All Tokens	+0.8	+10.6	+0.2	+1.5	+0.4	+1.0
All Non-Sink Tokens	+0.9	+19.5	+0.2	+1.9	+0.6	+1.2
All Visual Tokens	+1.3	<u>+30.1</u>	<u>+0.4</u>	<u>+2.6</u>	<u>+0.8</u>	<u>+1.9</u>
Query-Guided Visual Tokens	**+2.1**	+20.9	+0.2	+ 0.0	-0.1	+1.8
Visual Non-Sink (G-SAR)	<u>+1.9</u>	**+42.9**	**+0.8**	**+4.3**	**+1.4**	**+3.1**

Table 9: **Additional ablation on attention redistribution targets on LLaVA-1.5-7B.** All values denote performance improvements over the Greedy baseline, with higher values indicating better performance. Bold and underline denote the best and second-best results, respectively.

Redistribution	ΔGQA	ΔMME	ΔPOPE-A	ΔCHAIR-S	ΔCHAIR-I	ΔMathVista
All Tokens	+0.4	+10.8	+0.1	+1.0	+0.3	+0.8
All Non-Sink Tokens	+0.7	+19.8	+0.3	+2.0	+0.6	+1.7
All Visual Tokens	+1.1	+30.7	<u>+0.6</u>	<u>+2.6</u>	<u>+0.9</u>	+2.7
Query-Guided Visual Tokens	**+1.8**	**+50.2**	+0.3	-1.0	+0.0	**+5.7**
Visual Non-Sink (G-SAR)	<u>+1.6</u>	<u>+43.8</u>	**+1.3**	**+4.4**	**+1.5**	<u>+4.5</u>

As shown in Table 7, performance exhibits a non-monotonic dependence on retention strength. A weak constraint $(0.1, 5)$ allows greater modification of sink attention, but does not yield the strongest improvements on most benchmarks. The intermediate setting $(0.2, 10)$ achieves the most consistent gains across general visual understanding and hallucination metrics. Beyond this setting, further strengthening the retention constraint progressively limits the attention mass available for redistribution, accompanied by reduced gains on GQA, MME, and hallucination benchmarks.

MathVista shows a different trend, with the improvement increasing from $+1.8$ to $+5.0$ as the retention constraint becomes stronger. This observation suggests that reasoning-intensive tasks may benefit from retaining more sink attention, which is compatible with the structural influence of AS observed in Sec. 4.2. Overall, the results indicate that neither overly weak nor overly restrictive retention is uniformly optimal across tasks. The default setting $(0.2, 10)$ provides the most consistent balance between preserving sink attention and retaining sufficient capacity for visual redistribution.

B.5 Additional Redistribution Results

To further assess the generality of our redistribution strategy, we evaluate the same redistribution targets on InternVL3.5-8B and LLaVA-1.5-7B, as shown in Tables 8 and 9. Together with the Qwen3-VL-8B results in Table 5, a consistent pattern emerges across all three architectures.

Broad redistribution targets, such as all tokens or all non-sink tokens, yield relatively modest gains. Restricting redistribution to visual tokens generally improves performance more substantially. For example, on InternVL3.5-8B, All Visual Tokens improves MME by $+30.1$, compared with $+10.6$

for All Tokens, while on LLaVA-1.5-7B the corresponding gains are $+30.7$ versus $+10.8$. Query-guided visual redistribution can further benefit general visual understanding and reasoning, but its hallucination behavior is less stable. On LLaVA-1.5-7B, it achieves the largest gains on MME ($+50.2$) and MathVista ($+5.7$), yet degrades CHAIR-S by -1.0; a similar degradation is observed on Qwen3-VL-8B.

By contrast, redistribution to visual non-sink tokens provides the most consistent balance across capabilities. On InternVL3.5-8B, G-SAR achieves the strongest improvements on MME ($+42.9$), POPE-A ($+0.8$), CHAIR-S ($+4.3$), CHAIR-I ($+1.4$), and MathVista ($+3.1$). On LLaVA-1.5-7B, it likewise provides the strongest hallucination gains ($+1.3$ POPE-A, $+4.4$ CHAIR-S, and $+1.5$ CHAIR-I), while remaining competitive on MME ($+43.8$) and MathVista ($+4.5$). Combined with the Qwen3-VL-8B results, these findings support the design of G-SAR: released sink attention is most effective when redirected toward visual evidence while excluding dominant sink positions.