BEYOND VISUAL ATTENTION MAGNITUDE: VISUAL SINK REWEIGHTING FOR MITIGATING MULTIMODAL HALLUCINATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (MLLMs) have demonstrated strong vision-language capabilities, yet they remain prone to multimodal hallucination, where generated responses are inconsistent with visual content. Recent studies have identified the Visual Attention Sink phenomenon, showing that MLLMs may allocate excessive attention to visually uninformative or semantically irrelevant visual tokens. While existing work has analyzed sink tokens and begun to distinguish sink-prone attention heads, the relationship between visual attention magnitude and the structural quality of head-level visual attention remains insufficiently understood. As a result, high-visual-attention heads may be mistakenly regarded as effective visual heads even when they exhibit sink-like patterns, whereas low-visual-attention heads with meaningful visual structures may be overlooked. In this paper, we conduct a structure-aware head-level analysis of Visual Sink and show that visual attention magnitude alone is insufficient for identifying effective visual heads. Motivated by this observation, we propose Visual Sink Reweighting (VSR), a training-free method for mitigating multimodal hallucination. VSR constructs a prototypical Visual Sink pattern and introduces SinkScore to measure the structural similarity between each attention head and the prototype. Based on SinkScore, VSR suppresses sink-prone heads while enhancing non-sink heads whose visual attention structures deviate from the Visual Sink prototype during generation. Extensive experiments on multiple benchmarks and MLLMs demonstrate that VSR consistently reduces hallucination and improves visual grounding without additional training or parameter updates.

1 INTRODUCTION

In recent years, MLLMs (Chen et al., 2023; Li et al., 2023a; Liu et al., 2023; 2024b; Zhu et al., 2024)have achieved remarkable progress across a wide range of vision-language tasks, including visual question answering, image captioning, and multimodal reasoning. By aligning visual representations with large language models, MLLMs exhibit strong capabilities in cross-modal understanding and language generation. Despite these advances, multimodal hallucination (Gunjal et al., 2024; Li et al., 2023b; Liu et al., 2024a; Lovenia et al., 2024)remains a critical obstacle to their reliability, where models generate content that is inconsistent with the given image or insufficiently grounded in visual evidence.

Existing methods primarily mitigate hallucination by improving the use of visual evidence, including enhancing visual representations (Jiang et al., 2024; Li et al., 2025a; Seo et al., 2026), modifying decoding distributions (Huang et al., 2024; Leng et al., 2024), and intervening in internal visual signals during generation (Yin et al., 2025; Wang et al., 2025; Yin et al., 2026; Zhu et al., 2026). These methods generally attribute hallucination to insufficient visual evidence or excessive reliance on language priors. However, they mainly address whether visual information is strengthened, leaving less explored how visual attention is spatially distributed over image tokens.

This limitation is exemplified by recent findings on Visual Attention Sink (Kang et al., 2025; Luo et al., 2025), where MLLMs allocate disproportionate attention to visually uninformative or semantically irrelevant visual tokens, often at fixed spatial positions. However, prior work has primarily

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.

focused on identifying sink tokens and sparsifying sink-prone heads, leaving the relationship between a head's total visual-attention magnitude and its spatial attention pattern underexplored.

At the head level, visual relevance is commonly approximated by the visual attention ratio, i.e., the proportion of attention assigned to image tokens (Jiang et al., 2025; Yin et al., 2025; Che et al., 2025). This ratio-based view implicitly equates greater visual attention with stronger visual grounding.

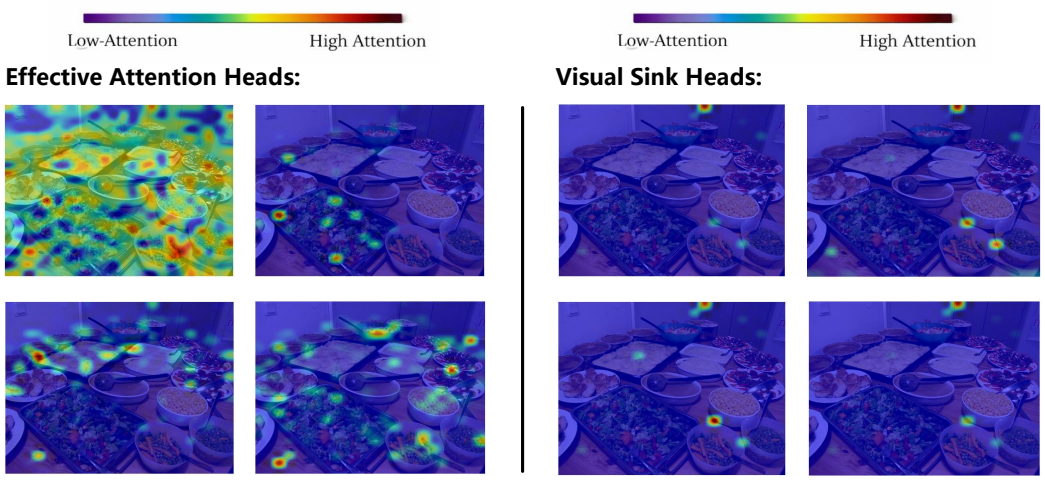


Figure 1: Visual attention maps of heads with different visual attention ratios. The visual attention ratios of the displayed heads are 90.5, 41.65, 3.77, 47.84, 39.8, 55.71, 51.67, and 5.01, in the order shown. High- and low-ratio heads both exhibit diverse attention structures.

Our analysis challenges this assumption. As shown in Figure 1, some high-ratio heads still collapse onto fixed and semantically irrelevant regions, exhibiting Visual Sink patterns, whereas some low-ratio heads preserve semantically meaningful spatial attention patterns. **This mismatch shows that visual attention magnitude and sink-like spatial patterns are not equivalent.** We therefore conduct a systematic comparison between visual attention magnitude and sink similarity. To quantify

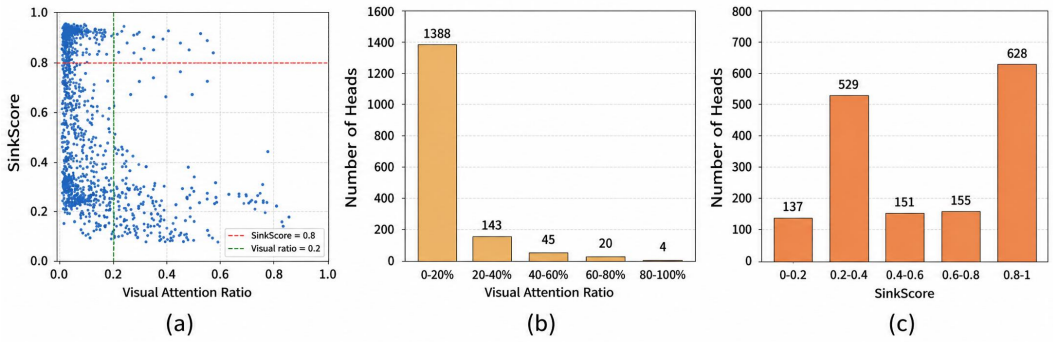


Figure 2: Quantitative motivation for structure-aware head characterization. (a) Relationship between visual attention ratio and SinkScore. (b) Distribution of visual attention ratios across heads. (c) Distribution of SinkScores across heads. Visual attention ratios are concentrated in a narrow range, whereas SinkScore provides a broader characterization according to structural similarity to the Visual Sink prototype.

this mismatch, we introduce SinkScore, which measures the similarity between each head's normalized visual attention map and a prototypical Visual Sink pattern. By normalizing the attention maps, SinkScore captures how attention is spatially distributed over image tokens, independently of the total visual attention mass. As shown in Figure 2(a), SinkScore exhibits no strict correspondence with the visual attention ratio, indicating that attention magnitude cannot reliably characterize

sink-like attention patterns. Figure 2(b) and 2(c) further show that visual attention ratios are heavily concentrated in the 0–20% range, whereas SinkScore produces a substantially broader distribution across heads. Consequently, ratio-based characterization provides limited resolution and may misidentify the heads that should be enhanced or suppressed. **SinkScore instead provides a spatial-pattern-based criterion for selecting intervention targets, allowing hallucination mitigation to distinguish sink-prone heads from those whose attention maps deviate from the Visual Sink prototype.**

Motivated by these findings, we propose **Visual Sink Reweighting (VSR)**, a training-free method that characterizes attention heads by their similarity to a Visual Sink prototype and selectively reweights their outputs to suppress sink-prone heads while enhancing non-sink heads, thereby mitigating multimodal hallucination. VSR first constructs a prototypical Visual Sink pattern from candidate attention maps with consistent spatial distributions and defines SinkScore as the similarity between each normalized head-level visual attention map and this prototype. Using two thresholds, VSR identifies high-SinkScore sink-prone heads, low-SinkScore non-sink heads, and heads with intermediate scores. During decoding, it down-weights sink-prone heads, up-weights non-sink heads, and leaves intermediate heads unchanged. Unlike token-level attention redistribution, VSR directly modulates head outputs. By guiding intervention with sink similarity rather than visual attention magnitude, VSR avoids reinforcing high-ratio sink-prone heads and overlooking low-ratio non-sink heads, thereby promoting more selective use of image-conditioned evidence and reducing image-inconsistent generation. The main contributions of this work are summarized as follows:

- **Head-level spatial-pattern analysis.** We present a head-level analysis of Visual Attention Sink and reveal a systematic mismatch between visual attention magnitude and spatial attention patterns: high-ratio heads may still exhibit sink-like patterns, whereas low-ratio heads may preserve semantically meaningful attention distributions. This finding shows that the visual attention ratio alone is insufficient for reliable head characterization.
- **Visual Sink Reweighting.** We propose VSR, a training-free method that constructs a Visual Sink prototype, introduces SinkScore to quantify head-level sink similarity, and conservatively reweights sink-prone, non-sink, and intermediate heads during decoding.
- **Effective hallucination mitigation.** VSR consistently reduces object hallucination and improves object- and attribute-level visual grounding without additional training or parameter updates.

2 Related Work

Hallucination in MLLMs. Multimodal Large Language Models (MLLMs) remain prone to hallucination, generating objects, attributes, or relations inconsistent with the visual input. Existing training-free methods mainly mitigate this issue through decoding-time correction or internal visual-signal enhancement. OPERA (Huang et al., 2024) penalizes overreliance on summary tokens and reallocates attention during decoding, while VCD (Leng et al., 2024) contrasts output distributions from original and distorted images to suppress unimodal priors. VISTA (Li et al., 2025b) reinforces visual information in the activation space and incorporates early-layer logits, whereas ClearSight (Yin et al., 2025)enhances visual signals in middle layers through Visual Amplification Fusion. Despite improving visual grounding from different perspectives, these methods do not explicitly model variations in head-level attention patterns, leaving how individual heads utilize visual information underexplored.

Visual Attention Sink in MLLMs. Recent studies have investigated visual attention sinks from complementary perspectives. At the decoder-token level, Kang et al. (Kang et al., 2025)trace visual sink tokens to massive activations in specific hidden-state dimensions and propose VAR, which focuses on token-level attention redistribution within selected heads. At the layer level, Zhang et al. (Zhang et al., 2025) reveal depth-dependent visual sink patterns and proposeEVAS, which propagates the attention map of a dense shallow-layer sink head to other heads within the same layer. From the perspective of sink origin and function, Luo et al. (Luo et al., 2025) study high-norm sink tokens emerging from the vision encoder and show that they can encode global visual semantics, while Choi et al. (Choi et al., 2026) distinguish ViT-emerged and LLM-emerged sinks and introduce layer-wise gating to balance global context and fine-grained visual evidence. These studies advance the understanding and manipulation of sink tokens, but they do not explicitly model

the mismatch between a head's total visual-attention magnitude and the sink-like spatial pattern of its attention map. In contrast, VSR constructs a prototypical Visual Sink pattern, characterizes each head through the similarity of its normalized visual attention map to this prototype, and directly reweights head outputs during inference.

3 METHOD

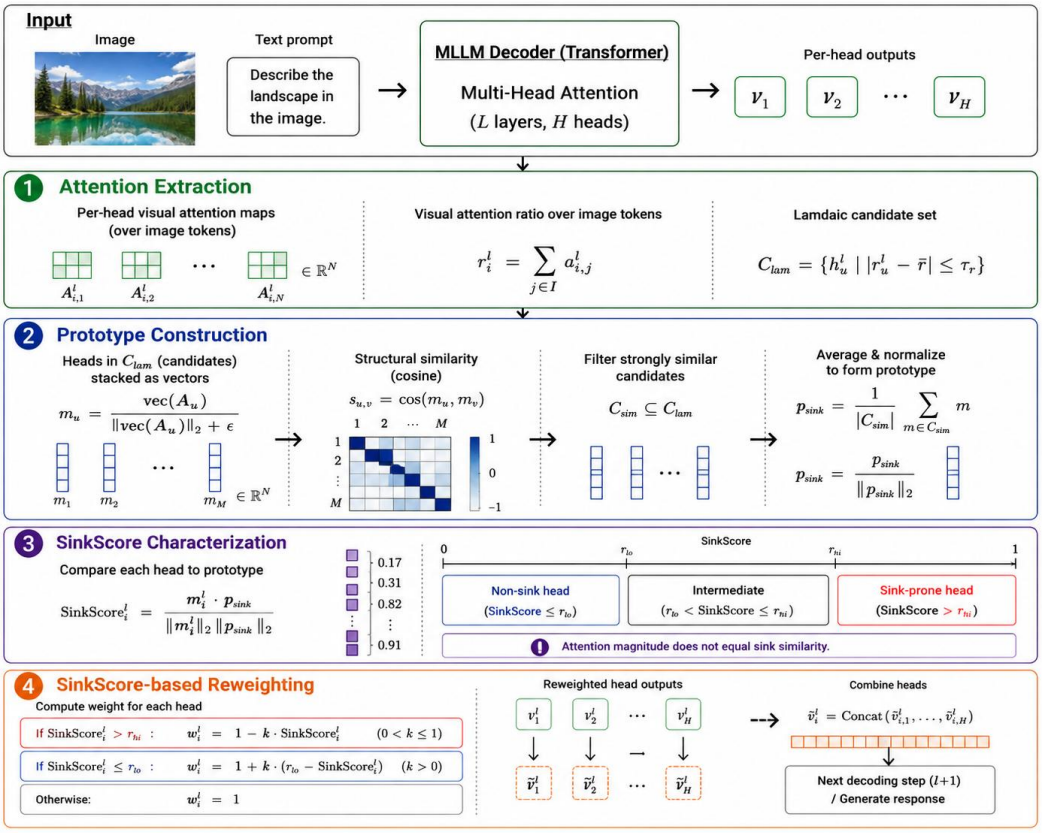


Figure 3: Framework of Visual Sink Reweighting (VSR). VSR constructs a Visual Sink prototype from structurally similar attention maps, measures head-level sink similarity with SinkScore, and performs SinkScore-guided head reweighting during inference.

In this section, we present **Visual Sink Reweighting (VSR)**, a training-free method consisting of three stages: Visual Sink prototype construction, SinkScore-based head characterization, and SinkScore-guided head reweighting. Given the normalized visual attention map of each head, VSR measures its similarity to the Visual Sink prototype, suppresses high- SinkScore sink-prone heads, enhances low-SinkScore non-sink heads, and leaves heads with intermediate scores unchanged. The overall framework is illustrated in Figure 3.

3.1 VISUAL SINK PROTOTYPE CONSTRUCTION

We consider an MLLM whose Transformer decoder consists of L layers, each containing H attention heads. During autoregressive generation, each attention head attends to both textual tokens and image tokens. Let I denote the set of image token indices. For head h in layer l at generation step t, we extract its attention weights over image tokens and denote the resulting visual attention map as $A^t_{l,h}$. The visual attention ratio of head h in layer l at step t is defined as the total attention mass assigned to image tokens:

$$r_{l,h}^{t} = \sum_{i \in \mathcal{I}} \alpha_{l,h}^{t,i}. \tag{1}$$

where $\alpha_{l,h}^{t,i}$ denotes the attention weight assigned to image token i by head h in layer l at generation step t. Our goal is to construct a Visual Sink prototype, defined as a representative normalized spatial distribution over image tokens that captures the common pattern shared by sink-like attention maps. We further observe that the spatial locations associated with Visual Sink are primarily determined by the input image rather than the textual context. Specifically, given the same input image, the visual tokens receiving excessive attention from sink-like heads remain highly consistent across different textual prompts and generation steps. This suggests that Visual Sink represents an image-dependent structural bias in the visual token space, rather than a transient attention pattern induced by specific language instructions. Therefore, attention maps exhibiting Visual Sink behavior across different decoding contexts can be aggregated to capture their shared spatial structure.We use low-ratio heads only to form a candidate pool, since low visual-attention magnitude does not necessarily imply a sink-like spatial pattern. Directly averaging all low-ratio maps would therefore mix sink and non-sink patterns, reducing the specificity of the resulting prototype.

Specifically, we select candidate visual attention maps according to the visual attention ratio:

$$\mathcal{C}_{\text{low}} = \left\{ A_{l,h}^{t} \mid r_{l,h}^{t} < \gamma \right\}. \tag{2}$$

where γ denotes the threshold for selecting low-ratio candidate maps. We then identify sink-like patterns within $\mathcal{C}_{\text{low}}$ according to the structural similarity among candidate visual attention maps. Since Visual Sink typically manifests as a stable attention pattern concentrated on fixed and semantically irrelevant visual positions, attention maps corresponding to Visual Sink are expected to be highly similar to each other. To compare attention structures while reducing the influence of attention magnitude, we vectorize each candidate attention map and apply ℓ_2 normalization:

$$m_{l,h}^{t} = \frac{\operatorname{vec}(A_{l,h}^{t})}{\| \operatorname{vec}(A_{l,h}^{t}) \|_2 + \epsilon}. \tag{3}$$

where vec$(\cdot)$ denotes vectorization and ϵ is a small constant for numerical stability. The normalized vector $m_{l,h}^{t}$ preserves the structural distribution of visual attention while reducing the effect of absolute attention magnitude. For any two normalized candidate maps mi and mj, we compute their structural similarity using cosine similarity:

$$s_{i,j} = \frac{m_i \cdot m_j}{\|m_i\|_2 \|m_j\|_2}. \tag{4}$$

Based on the pairwise similarities, we select a subset of candidate maps with strong internal similarity, denoted as $\mathcal{C}_{\text{sink}}$. This subset corresponds to attention maps that share a consistent sink-like structure. Compared with directly averaging all low-visual-attention maps, this procedure filters out non-sink attention patterns that have low visual attention ratios but do not exhibit Visual Sink behavior. Finally, we compute the centroid of the normalized attention maps in $\mathcal{C}_{\text{sink}}$ as the Visual Sink prototype:

$$P_{\text{sink}} = \frac{1}{|\mathcal{C}_{\text{sink}}|} \sum_{m \in \mathcal{C}_{\text{sink}}} m, \qquad P_{\text{sink}} \leftarrow \frac{P_{\text{sink}}}{\|P_{\text{sink}}\|_2 + \epsilon}. \tag{5}$$

The resulting P_{sink} captures the structural pattern of Visual Sink, where attention consistently concentrates on fixed and semantically irrelevant visual positions. We use this prototype to measure the structural similarity between each attention head and the Visual Sink pattern, which serves as the basis for defining SinkScore and characterizing attention heads according to their structural similarity to the Visual Sink prototype.

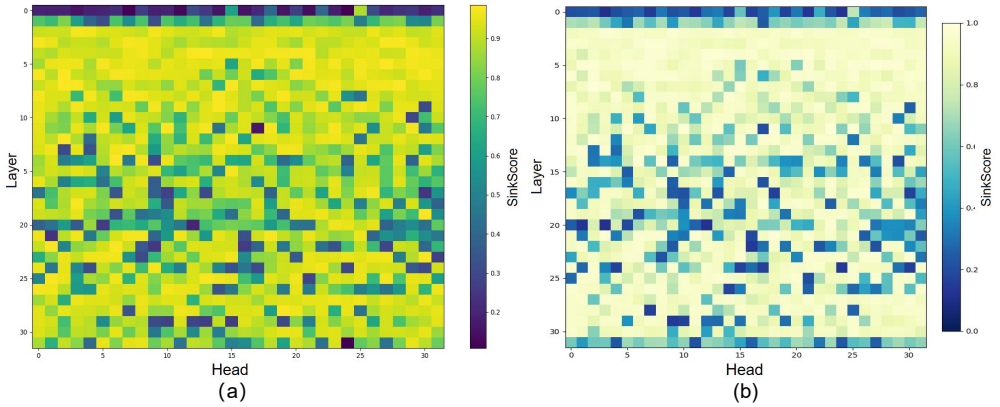


Figure 4: SinkScore distributions across Transformer layers and attention heads in (a) LLaVA-v1.5-7B and (b) LLaVA-v1.5-13B. The horizontal and vertical axes denote attention heads and Transformer layers, respectively. Darker colors indicate lower SinkScores, whereas lighter colors indicate higher SinkScores.

3.2 SinkScore for Head Characterization

After constructing the Visual Sink prototype, we define SinkScore to quantify the similarity between each head's normalized visual attention map and the prototype. For head h in layer l at generation step t, SinkScore is computed as

$$\mathrm{SinkScore}_{l,h}^{t} = \frac{m_{l,h}^{t} \cdot P_{\mathrm{sink}}}{\|m_{l,h}^{t}\|_2 \|P_{\mathrm{sink}}\|_2}. \tag{6}$$

A higher SinkScore indicates stronger similarity to the Visual Sink prototype and thus a greater tendency to exhibit sink-like attention; we refer to such heads as sink-prone heads. Conversely, a lower SinkScore indicates weaker similarity to the prototype, and such heads are termed non-sink heads. To avoid forcing heads with intermediate sink similarity into either category, VSR adopts two thresholds: τ_{sink} for identifying sink-prone heads and τ_{ns} for non-sink heads. We set $\tau_{\mathrm{sink}} = 0.8$ and $\tau_{\mathrm{ns}} = 0.6$ by default and analyze their sensitivity in Section 4.2. Each head is categorized as

$$\mathrm{HeadType}_{l,h}^{t} = \begin{cases} \text{Sink-Prone}, & \mathrm{SinkScore}_{l,h}^{t} > \tau_{\mathrm{sink}}, \\ \text{Non-Sink}, & \mathrm{SinkScore}_{l,h}^{t} < \tau_{\mathrm{ns}}, \\ \text{Intermediate}, & \text{otherwise}. \end{cases} \tag{7}$$

Figure 4 shows similar depth-dependent patterns across the two models. The first two layers are dominated by low- SinkScore heads, several subsequent early layers contain more high-SinkScore heads, and the middle layers exhibit substantial head-wise variation. This non-uniform distribution suggests that sink-like attention depends on both Transformer depth and individual heads, motivating head-specific rather than uniform layer-wise intervention. This characterization guides the head reweighting described next.

3.3 SinkScore-guided Head Reweighting

Based on SinkScore, VSR performs inference-time head reweighting to modulate the contribution of different attention heads during generation. Instead of redistributing attention among visual tokens within individual heads, VSR adjusts the contribution of each head according to its structural similarity to the Visual Sink prototype. Let o^t_l, h l,h denote the output of head h in layer l at generation step t. VSR assigns a scalar weight wt l,h to each head:

$$w_{l,h}^t = \begin{cases} 1 - b\,\mathrm{SinkScore}_{l,h}^t, & \mathrm{SinkScore}_{l,h}^t > \tau_{\mathrm{sink}}, \\ 1 + a(1 - \mathrm{SinkScore}_{l,h}^t), & \mathrm{SinkScore}_{l,h}^t < \tau_{\mathrm{ns}}, \\ 1, & \text{otherwise.} \end{cases} \tag{8}$$

$$\widetilde{o}_{l,h}^t = w_{l,h}^t \cdot o_{l,h}^t. \tag{9}$$

For heads with high SinkScores, VSR reduces their contributions because their attention patterns strongly resemble the Visual Sink prototype. For heads with low SinkScores, VSR increases their contributions because they deviate from sink-like patterns and are more likely to provide useful visual grounding. For heads with intermediate SinkScores, VSR leaves their contributions unchanged to avoid intervening in heads whose roles are ambiguous. Finally, the reweighted head outputs are combined following the standard multi-head attention computation:

$$\widetilde{o}_l^t = \mathrm{Concat}(\widetilde{o}_{l,1}^t, \widetilde{o}_{l,2}^t, \ldots, \widetilde{o}_{l,H}^t). \tag{10}$$

This design allows VSR to adaptively suppress sink-prone heads while enhancing non-sink heads during inference. Since VSR only modifies the contribution of attention heads and does not introduce additional trainable parameters or require model fine-tuning, it provides a lightweight and interpretable approach for improving visual grounding and mitigating multimodal hallucination.

4 Experiment

This section evaluates the effectiveness of the proposed VSR method in mitigating multimodal hallucination. We first introduce the experimental settings, including evaluation benchmarks, metrics, and backbone models. We then present quantitative results showing that VSR consistently reduces multimodal hallucination across multiple benchmarks and backbone models.

4.1 Experimental Settings

In Section 4.1, we introduce the evaluation benchmarks and metrics used in our experiments. In Section 4.1, we describe the MLLM backbones adopted for evaluation.

Datasets & Evaluation Metrics. CHAIR. We evaluate object hallucination in open-ended image captioning using CHAIR (Rohrbach et al., 2018). Following prior work (Huang et al., 2024; Leng et al., 2024; Li et al., 2025b; Yin et al., 2025), we use 500 randomly sampled images from the MSCOCO 2014 validation set and report sentence-level and instance-level hallucination rates, denoted as CHAIRS and CHAIRI, respectively. Lower values indicate fewer hallucinations. POPE. We evaluate object-existence hallucination using POPE (Li et al., 2023b) under the Random, Popular, and Adversarial settings. We report Accuracy and F1 score, where higher values indicate better object grounding. MME. We further evaluate VSR on MME (Fu et al., 2025), focusing on hallucination-related perception subsets, including object existence, counting, position, and color. These tasks assess object- and attribute-level visual grounding, with higher scores indicating better performance.

MLLM Backbones . We evaluate VSR on three representative open-source MLLMs: LLaVA-v1.5-7B, LLaVA-v1.5-13B (Liu et al., 2024b), and Qwen-VL-7B (Bai et al., 2023). LLaVA-v1.5-7B and LLaVA-v1.5-13B share a similar model design but differ in language-model scale, allowing us to examine whether VSR remains effective under different backbone capacities. Qwen-VL-7B belongs to a different MLLM family, which further evaluates the generality of VSR beyond a single architecture. Since VSR is a training-free inference-time method, all backbone models are kept frozen during evaluation, and no additional fine-tuning or external supervision is introduced.

4.2 Results and Analysis

Table 1 reports the CHAIR results on MSCOCO. Compared with regular decoding, VSR consistently reduces object hallucination on the evaluated backbones. The improvement is clear on

Table 1: CHAIR evaluation on MSCOCO. Lower sentence-level ($CHAIR_S$) and instance-level ($CHAIR_I$) hallucination rates are better.

	LLaVA-v1.5-7B		LLaVA-v1.5-13B		Qwen-VL-7B	
Method	S	I	S	I	S	I
Regular	45.0	14.7	20.6	6.2	46.0	12.5
VCD	47.4	13.0	21.7	6.5	45.2	12.3
VAR	43.2	13.8	17.3	5.1	–	–
VAF	45.6	13.3	21.7	5.9	45.6	12.1
VSR	40.6	12.9	16.2	4.3	41.3	11.3

Table 2: Performance on POPE. Accuracy and F1 are reported for Random, Popular, and Adversarial settings; higher is better.

		LLaVA-7B		LLaVA-13B		Qwen-VL-7B	
Category	Method	Acc.	F1	Acc.	F1	Acc.	F1
Random	Regular	81.2	80.5	81.3	80.2	87.2	86.7
	VCD	87.4	86.2	87.5	86.4	88.1	87.3
	VAF	87.9	88.4	88.1	87.9	88.4	87.9
	VSR	89.4	89.5	89.0	89.1	90.1	89.8
Popular	Regular	75.3	76.2	75.7	76.3	80.3	81.2
	VCD	82.2	82.9	82.7	83.1	81.1	82.2
	VAF	84.1	85.3	84.9	85.3	83.0	84.1
	VSR	86.2	86.8	85.9	86.8	84.1	85.2
Adversarial	Regular	73.2	74.4	73.8	74.5	74.2	74.9
	VCD	74.1	76.6	74.2	76.7	75.8	77.1
	VAF	76.7	79.9	76.1	79.1	75.9	77.1
	VSR	78.8	81.2	79.7	81.9	78.0	80.1

LLaVA-v1.5-7B and Qwen-VL-7B, where both sentence-level and instance-level hallucination rates decrease. On LLaVA-v1.5-13B, where the baseline hallucination rate is already relatively low, VSR still improves over regular decoding. These results indicate that SinkScore-guided head reweighting can reduce hallucinated object mentions in open-ended caption generation.

Table 2 shows the results on POPE under Random, Popular, and Adversarial settings. VSR achieves the best performance across all evaluated models and settings. The gains are especially evident under the Popular and Adversarial settings, where object queries are more likely to activate language priors and induce false positive predictions. These results suggest that VSR effectively improves object-existence grounding by suppressing heads whose attention patterns are structurally similar to the Visual Sink prototype.

Table 3 and Figure 5 report the results on hallucination-related subsets of MME. VSR improves the total score on all three backbones, demonstrating its effectiveness beyond object hallucination in captioning and binary probing. The improvements cover both object-level tasks, including existence and counting, and attribute-level tasks, including position and color. This indicates that VSR can enhance visual grounding in both object recognition and fine-grained visual perception.

Further experimental analyses, including hyperparameter sensitivity, head masking studies, and qualitative comparisons of different head selection strategies, are presented in the Appendix to provide additional evidence for the effectiveness and robustness of VSR.

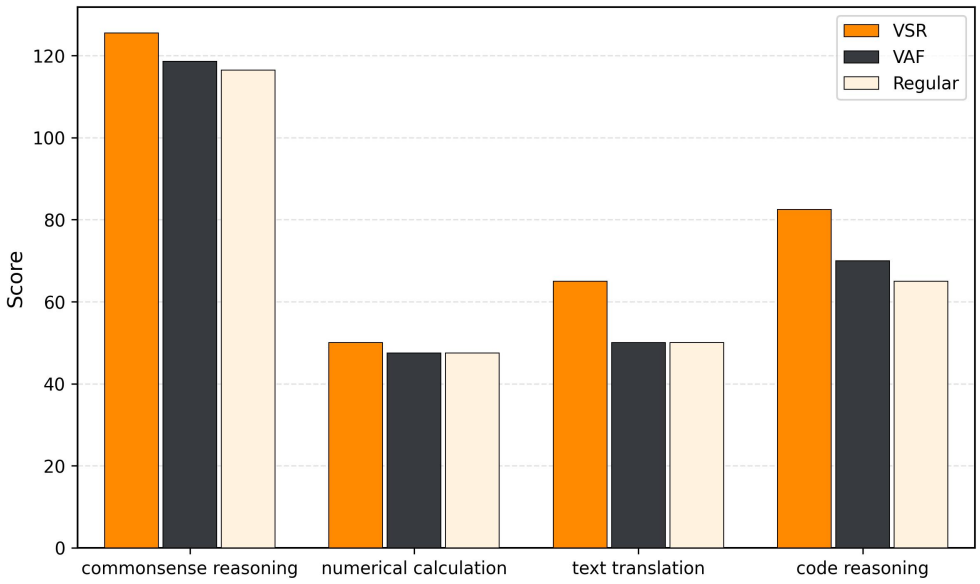


Figure 5: Performance on MME cognition tasks. We report the scores of different methods on cognition-oriented subsets of MME across three MLLM backbones. Higher scores indicate better cognitive reasoning performance.

Table 3: MME hallucination-related perception scores. Higher is better.

Model	Method	Existence	Count	Position	Color	Total
LLaVA-7B	Regular	185.00	146.67	128.33	150.00	610.00
	VCD	185.00	141.33	128.33	153.00	607.66
	VAF	185.00	151.67	133.33	155.00	625.00
	VSR	195.00	158.33	133.33	155.00	641.66
LLaVA-13B	Regular	185.00	155.00	133.33	165.00	638.33
	VCD	185.00	155.00	130.00	168.33	638.33
	VAF	183.33	153.33	131.67	165.00	633.33
	VSR	195.00	160.00	136.67	170.00	661.67
Qwen-VL-7B	Regular	158.33	150.00	128.33	170.00	606.66
	VCD	158.33	150.00	133.33	175.00	616.66
	VAF	128.33	151.67	128.33	170.00	578.33
	VSR	165.00	155.00	133.33	175.00	628.33

5 Conclusion

In this paper, we study multimodal hallucination from the perspective of head-level visual attention structures. We show that visual attention magnitude alone is insufficient for identifying effective visual heads, since some low-visual-attention heads still preserve meaningful visual grounding, while some high-visual-attention heads exhibit Visual Sink patterns. Based on this observation, we propose **Visual Sink Reweighting (VSR)**, a training-free inference-time method that constructs a Visual Sink prototype, defines SinkScore to characterize each attention head, and reweights head outputs accordingly. By suppressing high-SinkScore sink-prone heads and enhancing low- SinkScore non-sink heads, VSR improves visual grounding without additional training or parameter updates. Experiments on CHAIR, POPE, and MME across multiple MLLM backbones demonstrate the effectiveness of VSR in reducing multimodal hallucination.

AI USE STATEMENT

AI USE STATEMENT

In this work, we used generative AI tools for language refinement, grammar correction, and improving the clarity and readability of the manuscript. We also used AI-assisted tools for formatting assistance, including minor adjustments to LaTeX code, tables, and figures.

We have not used generative AI tools for research ideation, methodology design, mathematical derivations, experimental design, data collection, data generation, implementation of the proposed method, or interpretation of experimental results. All research decisions, technical contributions, experimental procedures, and conclusions were developed and verified by the authors.

Additionally, we used generative AI tools for recommended disclosure tasks related to improving the presentation quality of the manuscript. All AI-assisted modifications were carefully reviewed and revised by the authors. We verified the correctness of the manuscript content, including all claims, equations, experimental results, tables, and figures. We take full responsibility for the final content of this work, including any text or artifacts produced with the aid of generative AI.

REFERENCES

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-VL: A versatile vision-language model for understanding, localization, text reading, and beyond, 2023.

Liwei Che, Tony Qingze Liu, Jing Jia, Weiyi Qin, Ruixiang Tang, and Vladimir Pavlovic. Hallucinatory image tokens: A training-free eazy approach to detecting and mitigating object hallucinations in lvlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 21635–21644, 2025.

Keqin Chen, Zhao Zhang, Weili Zeng, Richong Zhang, Feng Zhu, and Rui Zhao. Shikra: Unleashing multimodal llm's referential dialogue magic, 2023. URL https://arxiv.org/abs/2306.15195.

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. *arXiv preprint arXiv:2604.03316*, 2026.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, Rongrong Ji, Caifeng Shan, and Ran He. MME: A comprehensive evaluation benchmark for multimodal large language models. In *Advances in Neural Information Processing Systems*, volume 38, 2025.

Anisha Gunjal, Jihan Yin, and Erhan Bas. Detecting and preventing hallucinations in large vision language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 18135–18143, 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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13418–13427, 2024.

Chaoya Jiang, Haiyang Xu, Mengfan Dong, Jiaxing Chen, Wei Ye, Ming Yan, Qinghao Ye, Ji Zhang, Fei Huang, and Shikun Zhang. Hallucination augmented contrastive learning for multimodal large language model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 27036–27046, 2024.

Zhangqi Jiang, Junkai Chen, Beier Zhu, Tingjin Luo, Yankun Shen, and Xu Yang. Devils in middle layers of large vision-language models: Interpreting, detecting and mitigating object hallucinations via attention lens. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 25004–25014, 2025.

Seil Kang, Jinyeong Kim, Junhyeok Kim, and Seong Jae Hwang. See what you are told: Visual attention sink in large multimodal models. *arXiv preprint arXiv:2503.03321*, 2025.

Sicong Leng, Hang Zhang, Guanzheng Chen, Xin Li, Shijian Lu, Chunyan Miao, and Lidong Bing. Mitigating object hallucinations in large vision-language models through visual contrastive decoding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13872–13882, 2024. doi: 10.1109/CVPR52733.2024.01316.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 19730–19742. PMLR, 2023a.

Wei Li, Zhen Huang, Houqiang Li, Le Lu, Yang Lu, Xinmei Tian, Xu Shen, and Jieping Ye. Visual evidence prompting mitigates hallucinations in large vision-language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4048–4080, 2025a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023b.

Zhuowei Li, Haizhou Shi, Yunhe Gao, Di Liu, Zhenting Wang, Yuxiao Chen, Ting Liu, Long Zhao, Hao Wang, and Dimitris N. Metaxas. The hidden life of tokens: Reducing hallucination of large vision-language models via visual information steering. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 35799–35819. PMLR, 2025b.

Fuxiao Liu, Kevin Lin, Linjie Li, Jianfeng Wang, Yaser Yacoob, and Lijuan Wang. Mitigating hallucination in large multi-modal models via robust instruction tuning. In *International Conference on Learning Representations*, volume 2024, pp. 57689–57733, 2024a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024b.

Holy Lovenia, Wenliang Dai, Samuel Cahyawijaya, Ziwei Ji, and Pascale Fung. Negative object presence evaluation (nope) to measure object hallucination in vision-language models. In *Proceedings of the 3rd Workshop on Advances in Language and Vision Research (ALVR)*, pp. 37–58, 2024.

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. *arXiv preprint arXiv:2510.08510*, 2025.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4035–4045, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1437.

Hoigi Seo, Dong Un Kang, Hyunjin Cho, Joohoon Lee, and Se Young Chun. On epistemic uncertainty of visual tokens for object hallucinations in large vision-language models. *Advances in Neural Information Processing Systems*, 38:130605–130660, 2026.

Sudong Wang, Yunjian Zhang, Yao Zhu, Enci Liu, Jianing Li, Yanwei Liu, and Xiangyang Ji. Shift: Smoothing hallucinations by information flow tuning for multimodal large language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 3639–3649, 2025.

Hao Yin, Guangzong Si, and Zilei Wang. Clearsight: Visual signal enhancement for object hallucination mitigation in multimodal large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14625–14634, 2025.

Jianghao Yin, Qin Chen, Kedi Chen, Jie Zhou, Xingjiao Wu, and Liang He. Dynamic multimodal activation steering for hallucination mitigation in large vision-language models. *arXiv preprint arXiv:2602.21704*, 2026.

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 *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 3512–3534, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.174.

Deyao Zhu, Xiaoqian Shen, Xiang Li, Mohamed Elhoseiny, et al. Minigpt-4: Enhancing vision-language understanding with advanced large language models. In *International Conference on Learning Representations*, volume 2024, pp. 18378–18394, 2024.

Xingyu Zhu, Kesen Zhao, Liang Yi, Shuo Wang, Zhicai Wang, Beier Zhu, and Hanwang Zhang. Look carefully: Adaptive visual reinforcements in multimodal large language models for hallucination mitigation. *arXiv preprint arXiv:2602.24041*, 2026.

A APPENDIX

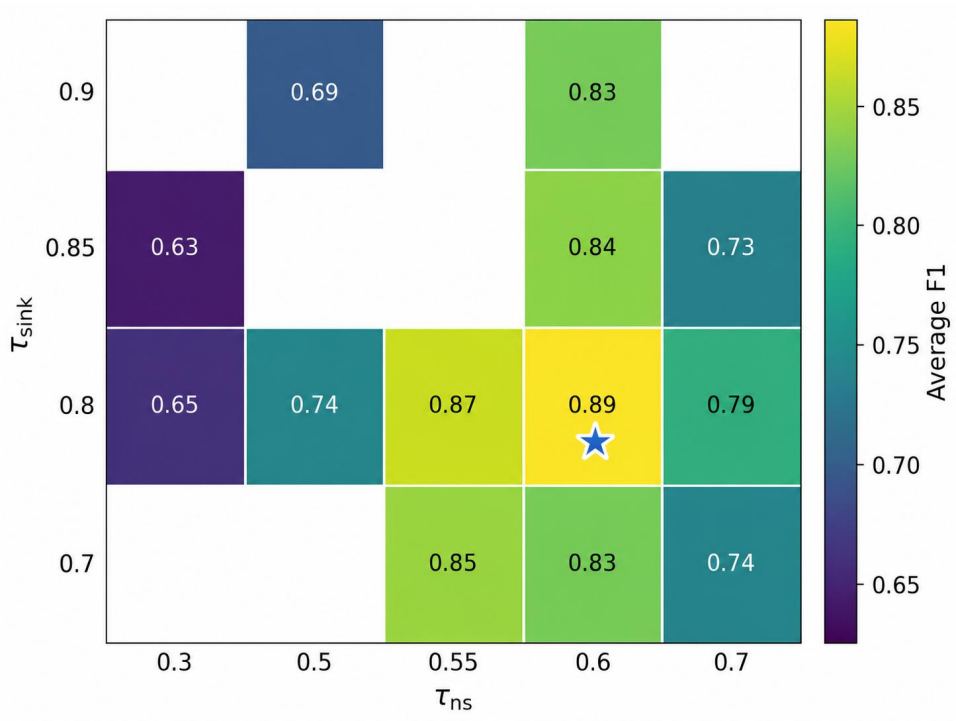


Figure 6: Two-dimensional sensitivity analysis of the SinkScore thresholds. We vary the non-sink threshold τ_{ns} and the sink-prone threshold τ_{sink}, and report the average F1 score under different threshold combinations. Blank cells denote configurations that were not evaluated. The star marks the default setting ($\tau_{\text{ns}} = 0.6$, $\tau_{\text{sink}} = 0.8$).

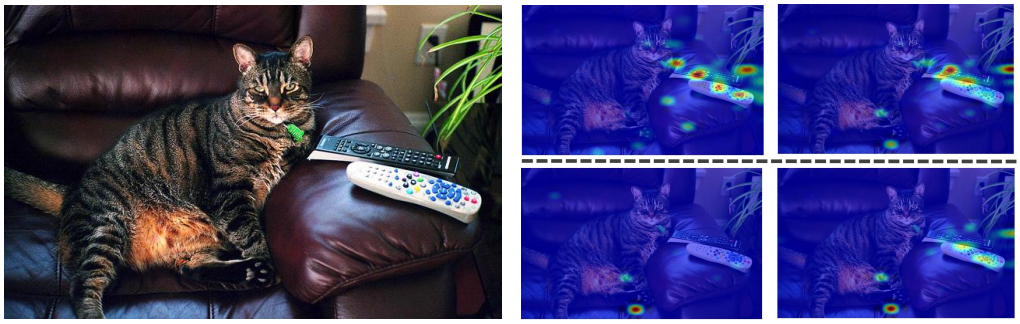
Hyperparameter Sensitivity. We further analyze the sensitivity of VSR to the two SinkScore thresholds, namely the non-sink threshold τ_{ns} and the sink-prone threshold τ_{sink}. Figure 6 shows that VSR achieves strong performance in a relatively stable region around (τ_{ns} = 0.55 $\sim$ 0.6, τ_{sink} = 0.8), indicating that the method is not overly sensitive to moderate variations of the threshold values. Based on this observation, we choose τ_{ns} = 0.6 and τ_{sink} = 0.8 as the default configuration in all experiments.

Head Selection and Masking Analysis. To further examine whether SinkScore identifies functionally distinct attention heads, we conduct both quantitative head masking analysis and qualitative head reweighting comparison. The masking analysis verifies whether the heads categorized by SinkScore play different roles in object grounding, while the qualitative comparison directly contrasts SinkScore-based head selection with the commonly used visual-attention-ratio-based criterion.

As shown in Table 4, masking sink-prone heads causes only marginal performance changes compared with the original model, suggesting that these heads contribute relatively limited effective visual evidence for object grounding under the evaluated settings. In contrast, masking non-sink heads leads to a substantial performance drop across different POPE settings and model backbones. For example, on LLaVA-v1.5-7B, Mask-NSH decreases the accuracy from 81.2 to 60.1 under the Random setting, and from 73.2 to 55.2 under the Adversarial setting. A similar degradation is observed on Qwen-VL-7B. These results provide functional evidence that the low-SinkScore non-sink heads are important for visual grounding, whereas the high-SinkScore sink-prone heads make a comparatively limited contribution to object grounding.

Table 4: Head masking analysis on POPE. Mask-SPH and Mask-NSH denote masking high-SinkScore sink-prone heads and low-SinkScore non-sink heads, respectively. The substantial degradation caused by Mask-NSH indicates the importance of low-SinkScore heads for object grounding.

		LLaVA-v1.5-7B		Qwen-VL-7B	
Category	Method	Acc.	F1	Acc.	F1
Random	Original	81.2	80.5	87.2	86.7
	Mask-SPH	81.2	80.7	87.1	86.7
	Mask-NSH	60.1	58.2	65.3	63.3
Popular	Original	75.3	76.2	80.3	81.2
	Mask-SPH	75.1	75.8	80.6	81.4
	Mask-NSH	56.3	58.4	65.5	66.7
Adversarial	Original	73.2	74.4	74.2	74.9
	Mask-SPH	71.2	72.5	74.1	74.6
	Mask-NSH	55.2	54.1	54.3	54.3



Orignal:The image features a large, fat cat lying on a brown leather couch, taking up a significant portion of the couch. The cat appears to be relaxed and comfortable, possibly enjoying a nap. **On the couch, there is a remote control** placed next to the cat, possibly indicating that the cat's owner was watching television or using a device before the cat decided to join them. The scene captures a cozy and peaceful moment between the cat and its environment.

VAF:The image features a large, fat cat lying on a brown leather couch. The cat is positioned in such a way that **it appears to be resting on top of a remote control**. The remote control is placed on the couch, close to the cat. The scene is quite relaxed and comfortable, with the cat enjoying its time on the couch.

VAR:The image features a large, fat cat lying on a brown leather couch. The cat is positioned on the couch, occupying a significant portion of the space. **A remote control is placed next to the cat, and a TV remote is also visible** on the couch. The scene appears to be a cozy and relaxed environment, with the cat enjoying its time on the couch.

Figure 7: Comparison of different head reweighting strategies.

We further provide a qualitative comparison to examine how different head selection criteria affect generation behavior. As shown in Figure 7, we compare the original model, VAF, and VSR on LLaVA-v1.5-7B. The four attention maps on the right illustrate the heads selected by the two intervention strategies.

Specifically, the upper two maps correspond to low-SinkScore non-sink heads enhanced by VSR, whereas the lower two maps correspond to high-visual-attention-ratio heads enhanced by VAF. This comparison highlights that the two methods select substantially different heads: VAF relies on the total amount of attention assigned to visual tokens, while VSR considers whether the spatial attention pattern resembles the Visual Sink prototype.

VAF still produces an unfaithful description, placing the cat on a remote control and missing the second remote control, showing that high visual-attention ratio is not a reliable proxy for visual grounding. In contrast, VSR correctly identifies two remote controls near the cat by enhancing low-SinkScore non-sink heads.

Consistent with the masking analysis, this result supports guiding head-level intervention by sink similarity rather than visual attention magnitude.