WinD: Mitigating Pruning-Induced Hallucinations in MLLMs via Window-Aware Diversity Selection

Anonymous authors
Paper under double-blind review

Abstract

Visual token pruning effectively accelerates Multimodal Large Language Models (MLLMs), yet its impact on model reliability remains under-explored. In this paper, we identify a critical side effect: prevalent pruning strategies significantly exacerbate hallucinations, particularly in complex multi-object scenes. Our analysis associates this degradation with the interaction between global CLS-guided selection and attention-sink behavior. Such selection disproportionately preserves sink-favored tokens, which occupy the limited preservation budget and reduce the grounding evidence available to the model, causing the model's focus to collapse onto a few sink tokens that cannot fully represent the whole visual context, while disregarding other essential visual information. To handle this problem, we shift the pruning paradigm from sink-seeking to preserving diverse important information, proposing the **Win**dow-aware **D**iversity selection (WinD) framework. WinD decentralizes the pruning process by adaptively partitioning the visual field into local windows based on visual content. It dynamically allocates pruning budgets based on local information density and introduces a semantic diversity-aware selection mechanism to suppress redundant sink features. Extensive experiments demonstrate that WinD strikes a superior balance between efficiency and reliability. On LLaVA-1.5-7B, our method reduces 88.9% of visual tokens and 77.5% of TFLOPs while retaining 96.7% of original performance, and effectively mitigates pruning-induced hallucinations. Our codes will be released.

1 Introduction

Multimodal Large Language Models (MLLMs) Bai et al. (2023); Liu et al. (2023); Li et al. (2023a); Tong et al. (2024) have demonstrated remarkable cross-modal perception and generation capabilities Chen et al. (2023a); Wu et al. (2024); Chen et al. (2024a) across diverse tasks such as visual question answering and instruction following Dai et al. (2023); Xu et al. (2023); Luo et al. (2023). Typically employing pre-trained visual encoders like CLIP Radford et al. (2021) or SigLIP Zhai et al. (2023), these models face substantial computational and memory overhead due to the massive number of visual tokens generated from high-resolution patches. As visual information is naturally sparser than text Achiam et al. (2023); Chen et al. (2023b); Devlin et al. (2019), visual token pruning Xing et al. (2024); Zhang et al. (2024); Yang et al. (2025); Chen et al. (2024b) has emerged as a dominant efficiency paradigm. By identifying and discarding redundant visual patches, these strategies achieve inference acceleration while maintaining performance.

However, we reveal that this acceleration comes at a significant, overlooked cost: **pruning-induced hallucination**. As illustrated in Figure 1, token pruning acts as a catalyst that exacerbates multimodal hallucinations, especially in complex, multi-object scenes. However, although we find this problem to be widespread in current work, few efforts have been made to analyze or handle it. Therefore, in this paper, we take a closer step towards this problem. Our analysis reveals that this phenomenon is not a mere byproduct of information loss, but a consequence of the **structural coupling** between pruning criteria and *attention sinks* Kang et al. (2025). Specifically, standard CLS-based pruning reinforces the model's inherent bias by collapsing the visual context onto a few "sink" tokens that lack representative semantic power, thereby depriving the model of the diverse evidence required for grounding.

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.

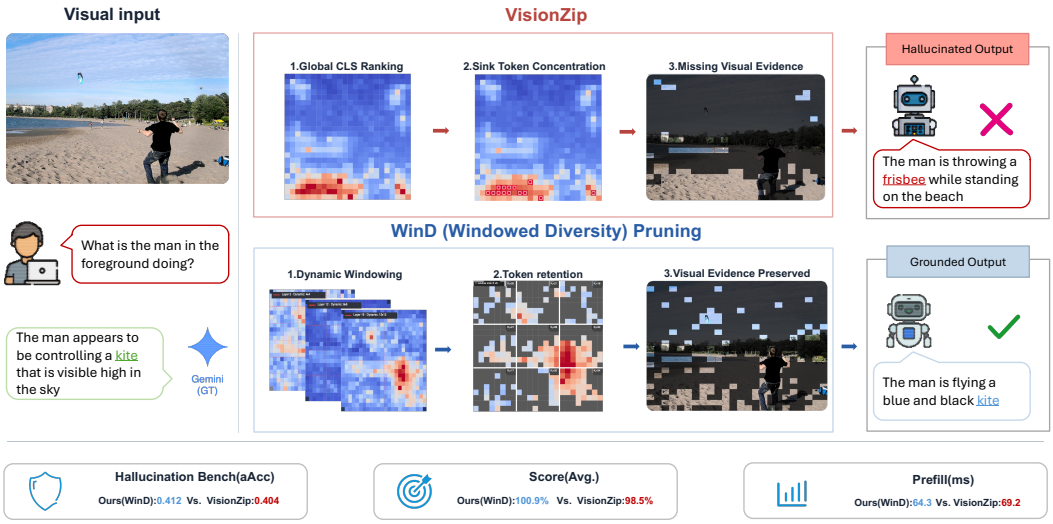


Figure 1: **Qualitative comparison of hallucination behaviors.** Global CLS Pruning methods may introduce hallucinated objects or attributes (highlighted in red). In this paper, we delve into this problem for an interpretation and solution.

To bridge this gap, we propose **WinD** (**Win**dow-aware **D**iversity Selection), a framework that transitions pruning from a global "sink-seeking" operation to a decentralized, content-adaptive process. Unlike prior entropy-based methods that only regulate interlayer budgets Tong et al. (2025), WinD leverages attention entropy to dynamically modulate spatial granularity and local diversity. Our framework comprises three synergistic components:

(1) Content-aware Dynamic Partitioning: Unlike current works Zhang et al. (2026); Zou et al. (2025) relying on **static, fixed-grid partitions**, we take visual contents into account to dynamically regulate both the global pruning budget and the spatial granularity of window partitioning. By shrinking window sizes in high-dispersion scenes, we force the model to look beyond global sinks and maintain a finer spatial grid for token selection.

(2) Diversity-balanced Quota Allocation: Unlike current token pruning works Yang et al. (2025) that sacrifice low-response regions, we prevent attention sinks from monopolizing the preservation budget by designing a balanced quota allocation strategy that guarantees representation for every spatial region, including low-response areas that may contain essential visual information. As illustrated in Figure 5, this ensures spatially diverse token retention even under aggressive compression.

(3) Semantic Diversity Selection: Within each window, we introduce a semantic penalty mechanism. By identifying local anchors and penalizing redundant features, we compel the model to retain a diverse set of informative tokens rather than solely focusing on a few important tokens.

Together, these mechanisms shift the pruning paradigm from *sink-seeking* to *preserving diverse important information*, suppressing the amplification of attention sinks and mitigating pruning-induced hallucinations in MLLMs.

In summary, the contributions are as follows:

- We quantify pruning-induced hallucination across benchmarks, showing that existing pruning strategies create a non-negligible problem in MLLMs.
- We trace pruning-induced hallucinations to global CLS-based selection, which amplifies the **Attention Sink** effect, reinforces inherent attention bias, and collapses focus onto a few tokens while discarding essential visual evidence.
- We propose **WinD**, replacing global sink-seeking with decentralized selection. Its content-aware dynamic windowing and semantic penalty balance spatial and semantic diversity, counteracting attention concentration.
- Extensive experiments show that WinD achieves a state-of-the-art balance between efficiency and reliability, reducing 77.5% of FLOPs while maintaining 96.7% performance and mitigating hallucination rates compared to existing pruning benchmarks.

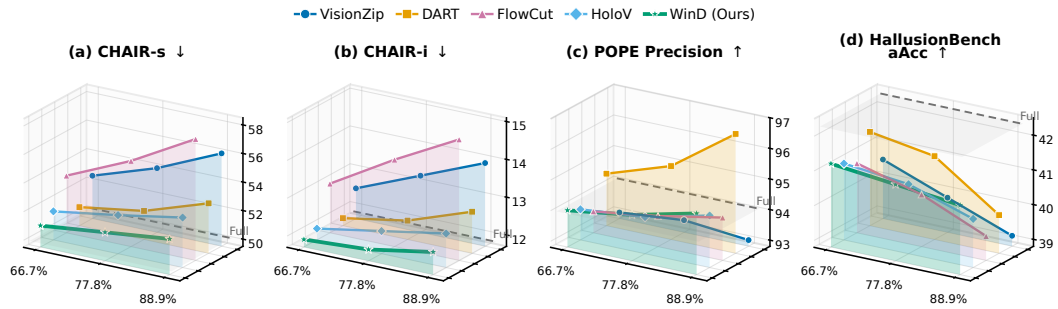


Figure 2: **Pruning-induced hallucination.** Dashed lines denote the unpruned baseline. CHAIR ↓; POPE precision and HallusionBench aAcc ↑.

2 Related Work

Visual token pruning is a primary technique for accelerating Multimodal Large Language Models (MLLMs). Existing research generally follows two paradigms: *attention-based* and *diversity-based* selection.

Attention-based methods prioritize high-saliency tokens. FastV Chen et al. (2024b), SparseVLM Zhang et al. (2024), and PyramidDrop Xing et al. (2024) use attention maps to identify and retain important visual regions. Although effective on simple tasks, these methods remain coupled to attention sinks, causing the model to over-focus on a few dominant tokens and lose information in complex multi-object scenes.

Diversity-based methods aim to reduce token redundancy. DART Wen et al. (2025) emphasizes suppressing duplicates, while DivPrune Alvar et al. (2025), CDPruner Zhang et al. (2025), and MMTok Dong et al. (2025) explore diversity maximization to maintain a holistic visual representation. Although these approaches avoid attention sink bias, they operate globally and lack spatial coverage guarantees, often neglecting critical evidence in low-response regions.

Both categories share a key limitation: global selection struggles to simultaneously maintain semantic diversity and spatial coverage. Spatial windowing alleviates this issue by partitioning visual tokens into local regions, as adopted by HoloV Zou et al. (2025) and VScan Zhang et al. (2026). Nevertheless, fixed-size partitioning cannot adapt to varying visual complexity and may introduce redundancy within local regions.

3 Analysis

In this section, we investigate the reliability degradation introduced by visual token pruning and uncover the underlying causes. We first quantitatively demonstrate that aggressive token pruning consistently amplifies multimodal hallucination across multiple benchmarks. We then analyze this phenomenon from the perspective of attention sinks, revealing how CLS-based pruning reinforces sink token dominance and causes the loss of grounding-critical visual evidence. Finally, we study the spatial effects of different pruning strategies and show that adaptive spatial selection is necessary to preserve both semantic diversity and spatial coverage.

3.1 Quantitative Evaluation of Pruning-induced Hallucination

Visual token pruning accelerates MLLM inference by reducing redundant visual tokens. However, we observe that aggressive compression compromises visual grounding and increases multimodal hallucination. To quantify this effect, we evaluate different pruning strategies under various compression ratios using three established benchmarks: CHAIR Rohrbach et al. (2018), POPE Li et al. (2023b), and HallusionBench Guan et al. (2024). For CHAIR, we construct a complex multi-object subset from COCO2014 Lin et al. (2014) and additionally report results on the full evaluation set in the appendix. Detailed evaluation protocols and metric definitions are provided in the appendix.

As shown in Figure 2, all evaluated pruning methods consistently degrade hallucination control across benchmarks. Notably, the degree of degradation varies significantly across methods: CLS-attention-based methods exhibit the most severe hallucination amplification, while diversity-based methods show markedly stronger robustness. These results demonstrate that hallucination amplification is not an inevitable consequence of token reduction itself, but is strongly influenced by the pruning criterion. The superior robustness of diversity-based methods suggests that *how* tokens are selected matters more than *how many* are retained. This observation motivates a deeper investigation into why CLS-attention-based pruning disproportionately amplifies hallucination compared to other strategies, which we address in the following subsection.

3.2 Hallucination Amplification under CLS-based Token Pruning: An Attention Sink Perspective

To understand why CLS-attention-based pruning disproportionately amplifies hallucination, we investigate prior studies on the causes of hallucinations in MLLMs. A recurring explanation in recent literature is the Attention Sink phenomenon, which has been identified as an important contributor to hallucination behaviors in large-scale Transformer models.

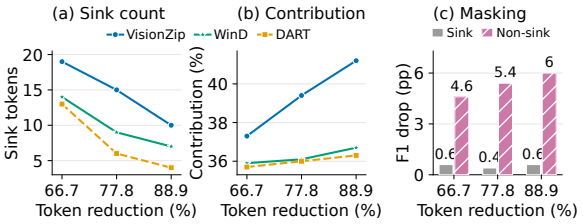


Figure 3: **Attention sink bias analysis.** (a) Retained sink tokens. (b) Their information contribution. (c) Causal intervention on retained tokens.

i) Attention Sink as a Source of Hallucination

Attention Sink refers to the tendency of models to aggregate global input information into a small number of tokens with limited semantic content, which then act as shortcuts for subsequent generation or classification Xiao et al. (2023); Ferrando & Voita (2024). This behavior is common in Transformer architectures; for example, Vision Transformers inject global information into special or background tokens to simplify downstream linear classification Dosovitskiy (2020); Darcet et al. (2023).

In multimodal settings, prior work suggests that excessive reliance on such tokens can cause the model to overlook complete visual evidence Huang et al. (2024). Wang et al. (2025) further show that when information aggregated by sink tokens is incomplete or biased, hallucinated content is more likely to be generated. At the visual level, Kang et al. (2025) observe that models consistently attend to a small set of image regions regardless of the input query, leading to the concept of Visual Attention Sink.

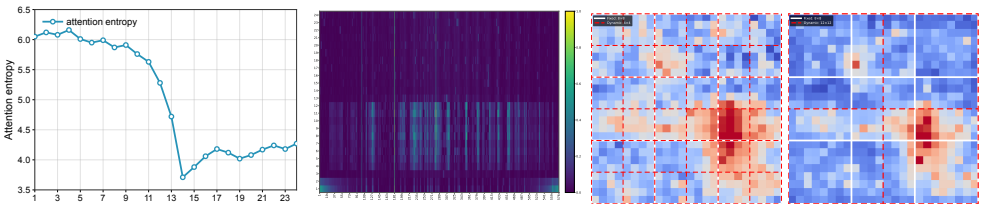
Notably, many existing token pruning methods rely on CLS token attention as the primary pruning criterion, preferentially removing tokens with low attention scores from the CLS token. This design closely aligns with the attention sink phenomenon, potentially reinforcing information concentration by preserving tokens favored by sink tokens while discarding others. Based on this observation, we raise the following hypothesis: ***Does CLS-based token pruning exacerbate hallucinations by amplifying attention sink bias?***

ii) CLS-based Pruning Over-Preserves Sink Tokens

To examine how different pruning criteria reshape the retained visual representation, we follow Kang et al. (2025) to detect sink tokens using the sink-dimension score $\varphi(\mathbf{x})$, defined as the maximum activation over a fixed set of sink dimensions $\mathcal{D}_{\text{sink}}$, normalized by the root mean square of the hidden state. We further quantify the contribution of each retained token to the residual stream as $\|\alpha_{i,j}^{l,h} x_j^{l-1} W_{OV}^{l,h}\|$.

As shown in Figures 3(a) and 3(b), VisionZip retains more sink tokens and assigns them a larger fraction of the total information contribution than DART. This indicates that unconstrained global CLS ranking disproportionately allocates the preservation budget to sink-favored tokens.

To move beyond correlational evidence, we causally intervene on the retained token set by masking an equal number of detected sink tokens or randomly sampled non-sink tokens. Figure 3(c) shows that removing non-sink tokens causes a substantially larger drop in POPE F1 across all compression



(a) (b) (c) (d)

Figure 4: **Layer-wise evolution of visual attention.** From left to right: CLS attention entropy across vision encoder layers, layer-wise CLS to visual token attention, and spatial attention maps at layers 5 and 19.

ratios, whereas masking sink tokens has only a marginal effect. This demonstrates that non-sink tokens carry substantially more grounding-critical visual evidence.

Together, these results support the following mechanism: global CLS-guided pruning over-preserves low-utility sink tokens; under a fixed token budget, these tokens crowd out grounding-critical non-sink evidence, and consequently amplify pruning-induced hallucination.

3.3 From Seeking Sink Tokens to Preserving Diverse Important Information

Beyond the attention-sink bias introduced by global token selection, we also need to specify a criterion for token selection. Therefore, we revisit the layer-wise evolution of the model attention to see the spatial variation of attention distributions.

i) Layer-wise Evolution of Attention Distribution

We first analyze how attention patterns evolve across the visual encoder. As shown in Figure 4(a-b), CLS attention exhibits distinct concentration patterns at different layers, with attention entropy varying significantly throughout the encoding process. These results indicate that a fixed pruning criterion may not capture the evolution of attention.

ii) Spatial Distribution of Different Pruning Strategies

The above observation indicates that visual information is spatially heterogeneous. In Fig. 3, we find that the spatial-diversity-based method (DART) shows good capability in preventing hallucinations. However, as the spatial heterogeneity of visual information indicates, its fixed partitions may not be optimal: although the spatial partitioning alleviates this issue by enforcing local coverage, fixed partitions cannot adapt to varying visual complexity. As shown in Figure 4(c-d), the fixed partitioning can hardly preserve informative regions with a fixed granularity given varying visual content.



Figure 5: Visualization of retained visual tokens under a fixed token budget. **Left:** single-object scene. **Right:** complex multi-object scene.

Moreover, Figure 5 visualizes the retained tokens under a fixed budget. Purple patches denote retained tokens, with darker colors indicating higher attention and information contribution, while red boxes mark detected sink tokens. In the single-object scene, VisionZip concentrates tokens around sink-dominated positions, while DART suppresses sink tokens but leaves parts of the subject uncovered. The differences become more pronounced in the multi-object scene. Fixed-window VScan allocates unnecessary quota to low-information sky regions and misses the small secondary traffic light, whereas the globally diversity-driven DART produces uneven spatial coverage.

These visualizations complement the intervention results: token selection needs not only to suppress sink concentration, but also preserve spatially distributed evidence. Therefore, in the following sections, we are inspired to design a dynamic-partitioning-based method to better balance hallucination prevention and performance preservation.

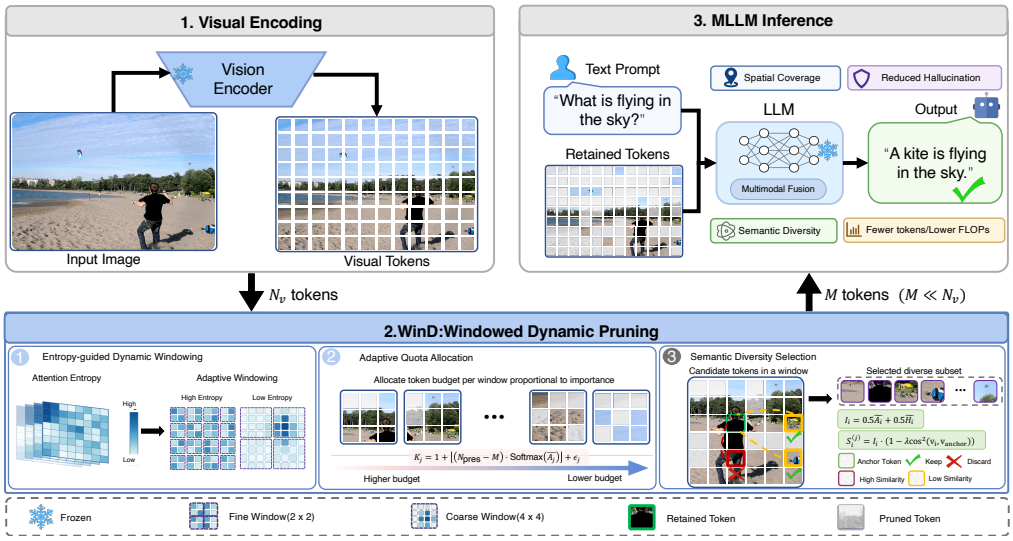


Figure 6: **Overview of the Window-aware Diversity Selection (WinD) framework.** WinD mitigates global attention bias through three stages: (1) **Entropy-aware partitioning** dynamically adjusts the window size S_l using global attention entropy r_H; (2) **Diversity-guaranteed quota allocation** distributes the budget N_{pres} with a base-plus-weighted strategy, ensuring $K_j \geq 1$ for every window; and (3) **Semantic selection** combines attention and feature entropy while penalizing similar features to preserve diverse, informative tokens.

4 METHOD

Based on the above analysis, we propose a dynamic token pruning framework to address the selection biases revealed in our empirical study. The core principle of our approach is to reduce over-reliance on CLS attention while preserving both semantic diversity and spatial coverage of visual tokens. The overall framework is illustrated in Figure 6.

4.1 PRELIMINARIES

In the Vision Encoder, a self-attention layer transforms the input sequence $\mathbf{X} \in \mathbb{R}^{N \times D}$ into a multi-head attention matrix $\mathbf{A} \in \mathbb{R}^{H \times N \times N}$, where H and N denote the head count and token number, respectively. We compute the head-average attention $\mathbf{A}_{avg} = \frac{1}{H}\sum_{h=1}^{H} \mathbf{A}^{(h)}$ and extract the global attention vector $\mathbf{a}_g = \mathbf{A}_{avg}[0, 1:] \in \mathbb{R}^{N-1}$, each element a_i reflects the contribution of token i to the CLS token. To evaluate attention sparsity, we calculate the normalized entropy:

$$r_H = \frac{-\sum_{i=1}^{N-1} a_i \log a_i}{\log(N-1)} \tag{1}$$

Following Tong et al. (2025), the number of tokens to be pruned, P, is determined by:

$$P = \text{round}\left(\frac{N - T_{target}}{\sqrt{L}} \cdot \sqrt{1 - r_H^2}\right) \tag{2}$$

where T_{target} is the final target token count and L is the remaining pruning layers. The preservation budget is thus $N_{\text{pres}} = N - P$.

4.2 ENTROPY-BASED DYNAMIC WINDOW PARTITIONING

To mitigate the reliance on global attention sinks, we adopt a strategy of dynamically partitioning the windows, which adaptively scales the local receptive field based on the spatial dispersion of visual features, ensuring that pruning preserves diverse visual evidence.

We use the normalized entropy r_H to quantify information dispersion and determine the target window size S_{val}:

$$S_{\text{val}} = S_{\max} - (S_{\max} - S_{\min}) \cdot r_H \tag{3}$$

The final window size S^l is selected from a set $\mathcal{S}_{cand} = \{S_{\min}, \ldots, S_{\max}\}$ by rounding S_{val} to the nearest candidate.

Unlike static windowing, we ensure an optimal trade-off between local detail and global context by **input-dependent dynamism** :

(i) For *high-dispersion* scenes (large r_H), we dynamically shrink the window size to enforce a finer spatial grid, preventing the omission of scattered visual evidence;

(ii) For *concentrated* scenes (small r_H), we expand the window to allow more flexible global selection.

This dynamic decentralization prevents the pruning process from prematurely collapsing into biased global shortcuts, thereby mitigating the vicious cycle of hallucination amplification.

4.3 Dynamic Quota Allocation

To prevent global attention sinks from monopolizing the budget N_{pres} and causing loss of localized visual evidence, we distribute quotas to windows W_j using a **base-plus-weighted** allocation strategy. We first define window importance as the mean CLS attention score $\bar{A}_j = \frac{1}{|W_j|} \sum_{i \in W_j} A_i$.

To ensure spatial diversity and mitigate hallucination-prone bias, the quota K_j for window W_j is determined by:

$$K_j = 1 + \lfloor (N_{\text{pres}} - M) \cdot \text{Softmax}(\bar{A}_j) \rfloor + \epsilon_j \tag{4}$$

where M is the number of windows, 1 is the mandatory base quota and ϵ_j is a residual bit greedily assigned by $\bar{A}_j$ to satisfy $\sum K_j = N_{\text{pres}}$. By mandating a base quota of one token per window, we explicitly decouple the pruning process from extreme attention concentration, ensuring that every spatial region contributes to the final representation. The remaining budget is distributed via Softmax-weighted importance, capped by window capacity. This balanced allocation preserves the fine-grained visual grounding necessary to break the "vicious cycle" of sink-induced hallucinations.

4.4 Diversity-Promoting Semantic Selection

Within each window W_j, we select K_j tokens that balance saliency and diversity. We define an importance score $I_i = 0.5\tilde{A}_i + 0.5\tilde{H}_i$, where $\tilde{H}_i$ is the normalized feature entropy. To suppress semantic redundancy, the token with the highest I_i is set as the *anchor*, and we penalize remaining tokens:

$$S_i^{(j)} = I_i \cdot \left(1 - \lambda \left| \frac{v_i \cdot v_{\text{anchor}}}{|v_i| \cdot |v_{\text{anchor}}|} \right|^2 \right) \tag{5}$$

The final set of retained tokens is $\mathcal{T} = \bigcup_j \text{Top-}K_j(S^{(j)})$. By explicitly penalizing semantic similarity, this mechanism forces the selection of unique features, maintaining the comprehensive grounding required to mitigate hallucinations.

5 Experiment

5.1 Experimental Settings

We evaluate LLaVA-1.5-7B Liu et al. (2023), LLaVA-NeXT-7B Liu et al. (2024), and Qwen2.5-VL Bai et al. (2025) on 12 multimodal understanding benchmarks under varying compression ratios. All pruning methods use identical checkpoints, prompts, token budgets, pruning schedules, and decoding configurations, following official implementations or reported settings. WinD uses one shared hyperparameter set across all benchmarks and backbones, without benchmark-specific tuning or method-specific prompting. Experiments run on NVIDIA RTX A6000 GPUs with 48 GB memory, with full hardware and software details in the appendix. We report one deterministic evaluation run per configuration.

Table 1: Performance comparison on LLaVA-1.5-7B across 12 multimodal benchmarks under extreme visual token compression.

Method	GQA	MMB	MMB^{CN}	MME	POPE	SQA	VQA^{V2}	VQA^{Text}	VizWiz	SEED	MM-Vet	LLaVA-B	Avg.
Upper Bound: 576 Tokens (100%)													
Vanilla	61.9	64.7	58.1	1862	85.9	69.5	78.5	58.2	50.0	59.3	31.1	66.9	100%
Retain 192 Tokens (66.7% Reduction)													
FastV (ECCV'24)	52.7	61.2	57.0	1612	64.8	67.3	67.1	52.5	50.8	57.1	27.7	49.4	89.4%
VisionZip (CVPR'25)	59.3	63.0	57.7	1778	85.2	68.7	76.6	57.3	51.2	56.4	31.7	67.7	98.5%
DART (EMNLP'25)	60.0	63.3	57.0	**1856**	82.8	**69.8**	76.7	57.4	51.2	56.9	30.9	67.6	98.7%
VScan (TMLR'26)	59.0	63.6	57.4	1806	86.2	68.6	76.9	57.7	50.4	57.1	32.4	67.9	99.0%
CDPruner (NeurIPS'25)	59.7	63.1	55.0	1821	86.2	69.0	76.6	56.2	51.2	57.5	33.8	67.5	99.1%
FlowCut (NeurIPS'25)	60.0	63.2	**57.8**	1836	86.1	68.6	77.1	57.5	**51.8**	57.5	34.3	68.4	100.2%
WinD(Ours)	**60.1**	**63.8**	57.5	1831	**86.9**	69.5	**77.1**	**57.6**	51.4	**57.7**	**36.7**	**68.9**	**100.9%**
Retain 128 Tokens (77.8% Reduction)													
FastV (ECCV'24)	49.6	56.1	56.4	1490	59.6	60.6	61.8	50.6	51.3	55.9	28.1	52.0	85.9%
VisionZip (CVPR'25)	57.6	62.0	56.2	1759	83.2	68.9	75.6	56.8	51.6	54.9	32.1	64.8	97.2%
DART (EMNLP'25)	58.7	**63.2**	56.9	**1840**	80.1	69.1	75.9	56.4	51.7	55.3	30.1	65.4	97.2%
VScan (TMLR'26)	58.4	63.0	56.9	1792	85.4	68.9	75.5	57.3	51.7	55.8	31.1	65.8	97.9%
CDPruner (NeurIPS'25)	57.5	61.1	53.2	1754	85.4	68.7	76.4	55.3	51.6	55.1	33.4	67.1	97.3%
FlowCut (NeurIPS'25)	58.5	62.1	56.5	1792	85.2	68.6	76.0	**57.3**	52.0	56.2	32.5	67.5	98.4%
WinD(Ours)	**58.8**	62.1	**57.0**	1761	**85.7**	**69.4**	**76.4**	56.7	**52.2**	**56.2**	**34.7**	**67.7**	**99.2%**
Retain 64 Tokens (88.9% Reduction)													
FastV (ECCV'24)	46.1	48.0	52.7	1256	48.0	51.1	55.0	47.8	50.8	51.9	25.8	46.1	77.5%
VisionZip (CVPR'25)	55.1	60.1	55.3	1687	77.0	69.0	72.4	55.5	52.8	52.2	31.5	62.9	94.4%
DART (EMNLP'25)	55.9	60.6	53.2	**1765**	73.9	69.8	72.4	54.4	51.6	52.0	29.8	63.4	93.7%
VScan (TMLR'26)	55.3	**62.1**	55.1	1698	82.4	69.1	72.8	55.4	51.8	52.7	31.1	64.2	95.3%
CDPruner (NeurIPS'25)	55.4	61.1	53.2	1735	82.1	69.2	72.6	55.3	53.4	52.4	31.7	64.3	95.4%
FlowCut (NeurIPS'25)	55.6	60.8	54.9	1744	81.9	69.1	72.8	**55.6**	52.9	53.2	32.3	65.1	96.0%
WinD(Ours)	**56.1**	60.7	**55.4**	1732	**82.9**	**69.8**	**73.2**	55.4	**53.4**	**53.5**	**33.4**	**65.3**	**96.7%**

5.2 RESULTS

Result on LLaVA. We evaluate WinD on **LLaVA-1.5-7B** and **LLaVA-NeXT-7B**. As shown in Tables 1 and 3, our framework achieves the highest average performance across various compression ratios. On LLaVA 1.5-7B, it retains 100.9%, 99.2% and 96.7% of the baseline performance with 192, 128 and 64 tokens, respectively. It also remains robust on the more complex LLaVA-NeXT-7B (initially 2880 tokens) under aggressive compression, demonstrating a strong efficiency–reliability trade-off across model scales.

Result on Qwen2.5-VL. As summarized in Table 4, WinD consistently outperforms FastV and HoloV across all compression tiers, and achieves competitive performance against FlowCut, surpassing it on the majority of benchmarks. At a 66.7% reduction, our method retains **98.6%** of vanilla performance, and maintains **90.9%** even under an aggressive 88.9% token reduction.

Efficiency analysis. We evaluate the computational efficiency of **WinD** on LLaVA-1.5-7B with a 64-token budget (88.9% reduction). As shown in Table 2, WinD achieves a significant 77.5% reduction in TFLOPs (8.82 $\rightarrow$ 1.98) compared to the vanilla baseline. While maintaining efficiency comparable to the state-of-the-art FlowCut (1.95 TFLOPs), WinD provides substantially higher reliability and reasoning performance, as demonstrated in previous benchmark results.

5.3 ANALYSIS AND DISCUSSION

The Precision–Recall Trade-off. Figure 7 reveals a clear trade-off on POPE: diversity-based methods achieve higher Precision but lower Recall, whereas global attention-based methods favor Recall at the cost of Precision, consistent with their stronger concentration on sink-favored tokens.

Table 2: Efficiency comparison on LLaVA-1.5-7B and LLaVA-NeXT-7B with 64 and 160 retained visual tokens, respectively. Bold and underlined metric values indicate the best and second-best results, respectively.

Method	Prefill (ms)↓	Speedup↑	TFLOPs↓	Red.↑
LLaVA-1.5-7B(576)	102.0	1.0×	8.82	–
+ FastV	87.3	1.1×	2.26	74.4%
+ VisionZip	69.2	1.4×	2.03	77.0%
+ CDPruner	65.4	1.5×	2.01	77.1%
+ FlowCut	**62.6**	**1.6**×	**1.95**	**77.9%**
+ WinD (Ours)	<u>64.3</u>	**1.6**×	<u>1.98</u>	<u>77.5%</u>

Method	Prefill (ms)↓	Speedup↑	TFLOPs↓	Red.↑
LLaVA-NeXT-7B(2880)	206.0	1.0×	31.03	–
+ FastV	117.0	1.8×	7.35	76.3%
+ VisionZip	69.9	2.9×	4.72	84.8%
+ CDPruner	66.7	3.1×	4.64	85.0%
+ FlowCut	**63.8**	**3.2**×	**4.29**	**86.2%**
+ WinD (Ours)	<u>65.3</u>	<u>3.2×</u>	<u>4.40</u>	<u>85.8%</u>

Table 3: Performance of LLaVA-NeXT-7B under different token-retention budgets.

Method	GQA	MMB	MMB^{CN}	MME	SQA	VQA^{V2}	Avg.
Upper Bound, 2880 Tokens (100%)							
Vanilla	64.2	67.9	60.6	1846	70.2	81.8	100%
Retain 640 Tokens (↓77.8%)							
PDrop	60.6	65.5	58.5	1781	–	78.3	96.0%
VisionZip	61.3	66.2	57.8	1787	68.7	79.1	96.6%
FlowCut	**61.9**	**66.7**	59.2	1837	68.8	79.8	97.9%
WinD (Ours)	61.3	66.5	**59.4**	**1844**	**70.0**	**80.1**	**98.2%**
Retain 320 Tokens (↓88.9%)							
PDrop	58.3	63.9	56.8	1736	–	75.2	92.9%
VisionZip	59.2	63.1	55.3	1702	68.1	76.2	93.1%
FlowCut	**59.8**	65.3	57.5	**1791**	68.4	76.8	95.4%
WinD (Ours)	59.6	**65.8**	**58.2**	1751	**69.4**	**77.2**	**95.6%**
Retain 160 Tokens (↓94.4%)							
PDrop	54.9	61.8	54.9	1513	–	70.2	87.0%
VisionZip	55.5	60.1	52.7	1628	68.3	71.4	89.1%
FlowCut	57.1	61.5	54.7	**1716**	68.1	74.6	91.8%
WinD (Ours)	**57.3**	**62.8**	**55.3**	1644	**68.7**	**74.9**	**91.9%**

Table 4: Performance of Qwen2.5-VL-7B under different token-retention budgets.

Method	GQA	MMB	MME	POPE	SQA	VQA^{Text}	Avg.
Upper Bound (100%)							
Vanilla	62.2	82.4	2327	86.9	84.7	84.9	100%
Token Reduction (↓66.7%)							
FastV	58.0	75.7	2130	82.1	78.5	77.9	93.5%
HoloV	59.7	77.9	2242	85.0	79.8	78.9	96.3%
FlowCut	**60.9**	79.2	2298	86.0	83.6	79.8	98.2%
WinD(Ours)	60.8	**79.4**	**2314**	**86.3**	**84.0**	**80.2**	**98.6%**
Token Reduction (↓77.8%)							
FastV	56.9	74.9	2036	80.7	78.0	72.0	90.8%
HoloV	58.3	76.1	2130	82.3	79.8	74.2	93.2%
FlowCut	**59.2**	76.3	2198	84.3	80.0	77.3	96.5%
WinD(Ours)	59.0	**78.1**	**2301**	**84.7**	**80.1**	**78.3**	**96.9%**
Token Reduction (↓88.9%)							
FastV	51.9	69.2	1962	78.6	77.4	60.3	84.1%
HoloV	55.7	72.4	2006	80.7	78.0	63.2	87.7%
FlowCut	**56.9**	72.1	2130	82.1	78.2	**68.9**	89.7%
WinD(Ours)	56.4	**72.8**	**2176**	**82.4**	**79.2**	68.7	**90.9%**

Table 5: Ablation study of WinD components on LLaVA-1.5-7B (128-token budget).

WinD Components			**Performance**		
DW	AQA	SDS	MME	GQA	POPE
			1709	56.5	82.1
✓			1731	58.5	84.3
✓	✓		1748	58.5	85.1
	✓	✓	1752	58.4	85.5
✓	✓	✓	**1761**	**58.8**	**85.7**

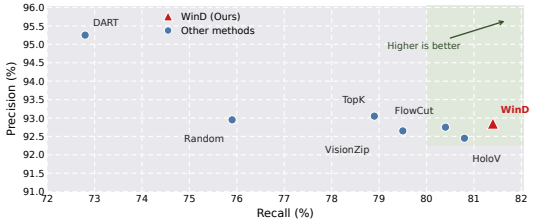


Figure 7: **Precision–Recall trade-off on POPE.** WinD achieves the best balance between precision and recall under a fixed token budget.

WinD bridges these regimes through adaptive windowing: smaller windows promote local diversity, while larger windows preserve global saliency. WinD achieves the highest recall while maintaining competitive precision, balancing hallucination control and visual coverage.

5.4 Ablation Studies

Effectiveness of WinD Modules. We investigate the individual and synergistic contributions of the components within the **WinD** framework: Dynamic Windowing **DW**, Adaptive Quota Allocation **AQA**, and Semantic Diversity Selection **SDS**. As summarized in Table 5, each module plays a distinct role in balancing performance and reliability. The full WinD configuration achieves the highest performance across all benchmarks, confirming that the integration of window-aware partitioning, adaptive allocation, and diversity-driven selection is key to maintaining a robust and efficient MLLM. Additional sensitivity experiments, including analyses of hyperparameter choices and partition strategies, are provided in the appendix.

6 Conclusion

In this paper, we identify a critical side effect of visual token pruning: amplified multimodal hallucinations. Our analysis shows that global CLS-guided pruning disproportionately retains low-utility sink tokens, crowding out grounding-critical non-sink evidence under constrained token budgets. To address this issue, we propose WinD, which preserves spatial coverage and semantic diversity through content-adaptive local selection. Extensive experiments across multiple benchmarks demonstrate that WinD achieves a strong balance between inference efficiency and model reliability while maintaining robust visual grounding, supporting the development of faster and more faithful MLLMs.

7 AI USE STATEMENT

This work used generative AI tools to assist with manuscript preparation, including improving the readability of the text, organizing tables and references, and drafting and revising portions of the text. We did not use generative AI tools for method design, experimental design, code implementation, or the analysis and interpretation of results, nor did we use AI to generate any reported numerical values. All AI-assisted text was reviewed and verified by the authors against the underlying code and experimental logs. We take responsibility for the final content of this work, including the text, the conclusions, and all associated artifacts.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9392–9401, 2025.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025.

Chi Chen, Ruoyu Qin, Fuwen Luo, Xiaoyue Mi, Peng Li, Maosong Sun, and Yang Liu. Position-enhanced visual instruction tuning for multimodal large language models. *arXiv preprint arXiv:2308.13437*, 2023a.

Fei-Long Chen, Du-Zhen Zhang, Ming-Lun Han, Xiu-Yi Chen, Jing Shi, Shuang Xu, and Bo Xu. Vlp: A survey on vision-language pre-training. *Machine Intelligence Research*, 20(1):38–56, 2023b.

Gongwei Chen, Leyang Shen, Rui Shao, Xiang Deng, and Liqiang Nie. Lion: Empowering multimodal large language model with dual-level visual knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26540–26550, 2024a.

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 *European Conference on Computer Vision*, pp. 19–35. Springer, 2024b.

Wenliang Dai, Junnan Li, Dongxu Li, Anthony Tiong, Junqi Zhao, Weisheng Wang, Boyang Li, Pascale N Fung, and Steven Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 36:49250–49267, 2023.

Timothée Darcet, Maxime Oquab, Julien Mairal, and Piotr Bojanowski. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pp. 4171–4186, 2019.

Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. Mmtok: Multimodal coverage maximization for efficient inference of vlms. *arXiv preprint arXiv:2508.18264*, 2025.

Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Javier Ferrando and Elena Voita. Information flow routes: Automatically interpreting language models at scale. *arXiv preprint arXiv:2403.00824*, 2024.

Tianrui Guan, Fuxiao Liu, Xiyang Wu, Ruiqi Xian, Zongxia Li, Xiaoyu Liu, Xijun Wang, Lichang Chen, Furong Huang, Yaser Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 14375–14385, 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.

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.

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 *International conference on machine learning*, pp. 19730–19742. PMLR, 2023a.

Yifan Li, Yifan Du, Kun Zhou, Jinpeng Wang, Wayne Xin Zhao, and Ji-Rong Wen. Evaluating object hallucination in large vision-language models. *arXiv preprint arXiv:2305.10355*, 2023b.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, January 2024. URL `https://llava-vl.github.io/blog/2024-01-30-llava-next/`.

Gen Luo, Yiyi Zhou, Tianhe Ren, Shengxin Chen, Xiaoshuai Sun, and Rongrong Ji. Cheap and quick: Efficient vision-language instruction tuning for large language models. *Advances in Neural Information Processing Systems*, 36:29615–29627, 2023.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Anna Rohrbach, Lisa Anne Hendricks, Kaylee Burns, Trevor Darrell, and Kate Saenko. Object hallucination in image captioning. *arXiv preprint arXiv:1809.02156*, 2018.

Jintao Tong, Wenwei Jin, Pengda Qin, Anqi Li, Yixiong Zou, Yuhong Li, Yuhua Li, and Ruixuan Li. Flowcut: Rethinking redundancy via information flow for efficient vision-language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL `https://openreview.net/forum?id=M6zQNbCaLl`.

Peter Tong, Ellis Brown, Penghao Wu, Sanghyun Woo, Adithya Jairam Vedagiri IYER, Sai Charitha Akula, Shusheng Yang, Jihan Yang, Manoj Middepogu, Ziteng Wang, et al. Cambrian-1: A fully open, vision-centric exploration of multimodal llms. *Advances in Neural Information Processing Systems*, 37:87310–87356, 2024.

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. *arXiv preprint arXiv:2501.15269*, 1, 2025.

Zichen Wen, Yifeng Gao, Shaobo Wang, Junyuan Zhang, Qintong Zhang, Weijia Li, Conghui He, and Linfeng Zhang. Stop looking for important tokens in multimodal language models: Duplication matters more. *arXiv preprint arXiv:2502.11494*, 2025.

Jiannan Wu, Muyan Zhong, Sen Xing, Zeqiang Lai, Zhaoyang Liu, Zhe Chen, Wenhai Wang, Xizhou Zhu, Lewei Lu, Tong Lu, et al. Visionllm v2: An end-to-end generalist multimodal large language model for hundreds of vision-language tasks. *Advances in Neural Information Processing Systems*, 37:69925–69975, 2024.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024.

Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11445–11465, 2023.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 19792–19802, 2025.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11975–11986, 2023.

Ce Zhang, Kaixin Ma, Tianqing Fang, Wenhao Yu, Hongming Zhang, Zhisong Zhang, Haitao Mi, and Dong Yu. Vscan: Rethinking visual token reduction for efficient large vision-language models. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=KZYhyilFnt.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *arXiv preprint arXiv:2506.10967*, 2025.

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. *arXiv preprint arXiv:2410.04417*, 2024.

Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. Don't just chase "highlighted tokens" in MLLMs: Revisiting visual holistic context retention. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=zKoeRtye8o.