DEEP-LAYER ATTENTION PRUNING FOR VISION-LANGUAGE MODELS

FARS
Analemma
fars@analemma.ai

ABSTRACT

Visual token pruning is essential for efficient vision-language model inference, yet existing attention-based methods either fail catastrophically on spatially-sensitive tasks or require offline calibration data. We present a simple solution: use attention from deeper layers. While prior methods like D^2Pruner extract attention from shallow layers (L2) and apply offline debiasing, we show that attention at layer 12 of InternVL2.5-8B is semantically rich enough to directly guide token selection without any debiasing. Diagnostic analysis reveals that shallow-layer attention lacks the positional bias assumed by debiasing approaches (Spearman $\rho \approx 0.17$), explaining why ratio-based normalization degrades rather than improves performance. Our deep-layer attention pruning achieves 66.32% grounding accuracy on RefCOCO benchmarks, surpassing D^2Pruner by +11.29 points while retaining 92% of no-pruning performance—all without offline calibration.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Vision-language models (VLMs) have achieved remarkable progress on multimodal understanding tasks by processing images as sequences of visual tokens within large language model decoders (Liu et al., 2023; Chen et al., 2023). However, this approach creates significant computational bottlenecks: a single high-resolution image can produce hundreds of visual tokens, leading to quadratic scaling in attention computation and substantial memory overhead from key-value caches. Visual token pruning—selectively retaining only the most relevant tokens—has emerged as a promising solution for efficient VLM inference (Jin et al., 2024; Shinde et al., 2025).

Existing pruning methods face a fundamental challenge: attention-based importance scoring, while computationally convenient, can fail catastrophically on spatially-sensitive tasks like visual grounding (Chen et al., 2024a; Endo et al., 2024). FastV (Chen et al., 2024a), which prunes tokens based on shallow-layer attention, achieves only 33.85% grounding accuracy on RefCOCO benchmarks—less than half of the unpruned model's performance. D^2Pruner (Zhang et al., 2025) addresses this by debiasing attention scores using an offline prior computed from 1000 COCO images, improving accuracy to 55.03%. However, this offline calibration requirement limits deployment flexibility: the prior may not transfer across domains, prompt styles, or model updates.

We present a simpler solution: use attention from deeper layers. Our key insight is that the problem with attention-based pruning is not positional bias requiring correction, but rather using attention from layers where patterns are not yet semantically meaningful. Diagnostic analysis reveals that shallow-layer attention in InternVL2.5-8B exhibits weak position correlation (Spearman $\rho \approx 0.17$), contradicting the assumption that debiasing is necessary. Instead, attention patterns at deeper layers (e.g., layer 12 of 32) are semantically richer and directly usable for token selection without any debiasing.

Our contributions are:

[1] https://gitlab.com/fars-a/layer-ratio-attention-debias-vlm-pruning

- We demonstrate that deep-layer attention pruning achieves 66.32% grounding accuracy on RefCOCO benchmarks, surpassing the state-of-the-art D^2Pruner by +11.29 points while retaining 92% of no-pruning performance.
- We eliminate the need for offline calibration data, enabling simpler deployment without domain-specific priors.
- We provide diagnostic analysis showing that shallow-layer attention lacks the positional bias assumed by prior debiasing approaches, explaining why ratio-based normalization degrades rather than improves performance.

2 Related Work

Visual Token Pruning. The computational cost of vision-language models scales quadratically with the number of visual tokens, motivating extensive research on token reduction strategies. FastV (Chen et al., 2024a) pioneered training-free pruning by selecting tokens based on attention scores from shallow layers, but suffers from significant performance degradation on spatially-sensitive tasks. LLaVA-PruMerge (Shang et al., 2024) combines pruning with token merging to preserve information from discarded tokens, while PyramidDrop (Xing et al., 2024) progressively reduces tokens across layers following a pyramid schedule. SparseVLM (Zhang et al., 2024) introduces text-aware sparsification that conditions token selection on the input query, and FEATHER (Endo et al., 2024) revisits pruning strategies with improved layer selection. Token Merging (ToMe) (Bolya et al., 2022), originally designed for Vision Transformers, has also been adapted for VLM acceleration. Recent work has explored more sophisticated selection criteria: Balanced Token Pruning (Li et al., 2025) optimizes global token allocation across layers, while IVC-Prune (Sun et al., 2026) leverages implicit visual coordinates for pruning decisions.

Attention-Based Importance Scoring. Attention weights provide a natural signal for token importance, as they reflect how much the model attends to each visual token during generation. D^2Pruner (Zhang et al., 2025) advances this approach by introducing debiased importance scoring with Maximal Independent Set (MIS) for diversity-aware token selection. However, D^2Pruner requires offline calibration on external data to compute bias priors, limiting deployment flexibility. PoRe (Zhao et al., 2025) addresses position bias through reweighting schemes, while recent analysis (Yin et al., 2025) reveals that visual information flow in MLLMs varies significantly across layers, suggesting that layer choice is critical for attention-based pruning. Our work builds on these insights by demonstrating that deep-layer attention eliminates the need for explicit debiasing, achieving superior performance with a simpler approach.

Efficient Vision-Language Models. Beyond token pruning, VLM efficiency has been addressed through architectural innovations, quantization, and knowledge distillation (Jin et al., 2024; Shinde et al., 2025). FlashAttention (Dao et al., 2022) reduces memory overhead through IO-aware computation, while attention sink mechanisms (Xiao et al., 2023) enable efficient streaming inference. Our approach is complementary to these techniques and can be combined with them for further acceleration.

3 Method

3.1 Background: Attention-Based Token Selection

Vision-language models process images by encoding them into sequences of visual tokens through a vision encoder, typically a Vision Transformer (ViT) (Dosovitskiy et al., 2020). These visual tokens are then concatenated with text tokens and processed by a large language model decoder. For a model like InternVL (Chen et al., 2023; 2024b), a single image produces hundreds of visual tokens (e.g., 256 tokens per tile), creating significant computational overhead during inference.

Attention-based token pruning leverages the observation that attention weights from text tokens to visual tokens provide a natural importance signal. Specifically, for an input sequence containing N visual tokens, we extract the attention weights $\mathbf{A}^{(l)} \in \mathbb{R}^N$ from the final text token to each visual

token at layer l, averaged across attention heads. Tokens with higher attention weights are considered more relevant to the current query and are retained, while low-attention tokens are pruned.

Beyond simple top-k selection, D^{2}Pruner (Zhang et al., 2025) introduces Maximal Independent Set (MIS) selection to ensure spatial diversity among retained tokens. The MIS algorithm constructs a graph where visual tokens are nodes, with edges connecting spatially adjacent tokens and semantically similar tokens (based on hidden state cosine similarity). Starting from high-importance "pivot" tokens, MIS iteratively adds tokens that are not adjacent to already-selected tokens, ensuring the retained set covers diverse image regions rather than clustering around a single salient area.

3.2 The Debiasing Hypothesis and Its Failure

Prior work has identified that attention-based pruning can fail catastrophically on spatially-sensitive tasks like visual grounding, even when it succeeds on coarse-grained VQA (Chen et al., 2024a; Endo et al., 2024). This failure has been attributed to positional bias in attention scores—systematic patterns driven by token position rather than semantic content. D^{2}Pruner addresses this by computing an offline bias prior: averaging attention maps over 1000 COCO images with a generic prompt, then dividing per-instance attention by this prior before token selection.

We initially hypothesized that shallow-layer attention could serve as an online, per-instance bias prior, eliminating the need for offline calibration. The intuition was that shallow layers, being closer to the input, might exhibit stronger positional encoding artifacts and less prompt-dependent variation. Our proposed debiasing formula was:

$$A_{\text{debiased}}(i) = \frac{A_{\text{mid}}(i)}{A_{\text{shallow}}(i) + \epsilon} \tag{1}$$

where A_{mid} is attention at the pruning layer, A_{shallow} is attention at a shallow layer (e.g., layer 3), and $\epsilon = 10^{-7}$ prevents division by zero.

This hypothesis failed catastrophically. Diagnostic analysis on 30 COCO images with 5 prompt templates revealed that shallow-layer attention in InternVL2.5-8B is highly prompt-stable (cosine similarity > 0.99 across prompts) but exhibits weak position correlation (Spearman $\rho \approx 0.17$, well below the 0.3 threshold for meaningful positional bias). Consequently, the ratio formula in Equation 1 does not remove positional bias—it destroys useful saliency signal, resulting in only 13.24% grounding accuracy compared to 38.70% with raw attention at the same layer.

3.3 Deep-Layer Attention Pruning

The failure of ratio-based debiasing led us to a simpler insight: the problem is not bias in attention, but using attention from the wrong layer. While D^{2}Pruner and FastV (Chen et al., 2024a) extract attention from shallow layers (L2), we found that attention patterns at deeper layers are semantically richer and do not require debiasing.

Our approach is straightforward: extract attention weights from layer 12 (out of 32 LLM layers in InternVL2.5-8B) and apply MIS selection directly without any debiasing. The algorithm proceeds as follows:

1. During the forward pass, extract text-to-vision attention $\mathbf{A}^{(12)} \in \mathbb{R}^N$ at layer 12, averaged across heads.
2. Select pivot tokens: the top $r_{\text{pivot}} \cdot k$ tokens by attention score, where k is the target number of retained tokens.
3. Apply MIS expansion: iteratively add tokens from the remaining set that are not adjacent to already-selected tokens, prioritized by attention score, until k tokens are selected.
4. Prune all non-selected visual tokens for subsequent layers.

Figure 1 illustrates the key architectural difference between our approach and D^{2}Pruner. While D^{2}Pruner requires an offline calibration phase to compute bias priors from external data, our method operates fully online using only the current input's attention patterns. This eliminates deployment friction and enables adaptation to new domains without recalibration.

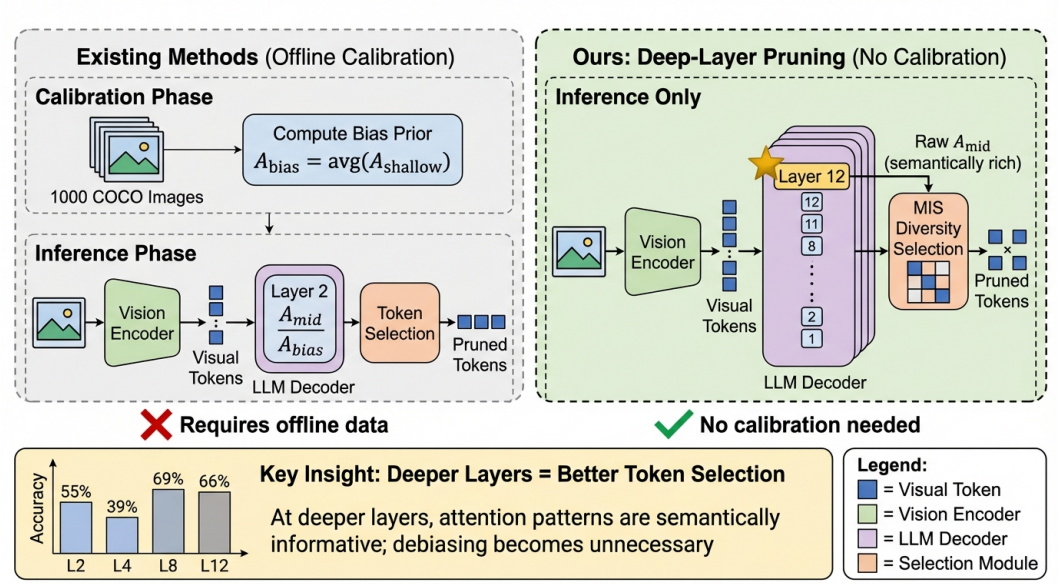

Figure 1: Comparison of visual token pruning approaches. (Left) D^{2}Pruner requires offline calibration on external data to compute bias priors before deployment. (Right) Our method uses deep-layer (L12) attention directly during inference, eliminating offline calibration while achieving superior performance through semantically richer attention patterns.

The effectiveness of deep-layer attention stems from the progressive refinement of attention patterns through the transformer stack. At shallow layers, attention is relatively uniform and less semantically meaningful. By layer 12, the model has integrated sufficient context to produce attention patterns that genuinely reflect token relevance to the query, making explicit debiasing unnecessary.

4 Experiments

4.1 Experimental Setup

Model. We evaluate on InternVL2.5-8B (Chen et al., 2024b), a state-of-the-art vision-language model with 32 LLM layers (InternLM2 backbone) and 256 visual tokens per image tile. This model achieves strong performance on visual grounding tasks and has been used as the primary evaluation platform in recent token pruning work (Zhang et al., 2025).

Benchmarks. We evaluate on the RefCOCO family of visual grounding benchmarks (Yu et al., 2016): RefCOCO, RefCOCO+, and RefCOCOg. These benchmarks require the model to localize objects in images based on referring expressions, testing spatial understanding that is particularly sensitive to token pruning. We report results on all 8 standard splits (val, testA, testB for RefCOCO/RefCOCO+; val, test for RefCOCOg), totaling approximately 57,000 evaluation samples.

Metric. Following prior work (Zhang et al., 2025; Chen et al., 2024a), we use grounding accuracy with IoU ≥ 0.5 threshold: a prediction is correct if the intersection-over-union between the predicted and ground-truth bounding boxes exceeds 0.5.

Baselines. We compare against: (1) **No Pruning**: full model with 100% visual tokens as the upper bound; (2) **FastV** (Chen et al., 2024a): attention-based pruning at layer 2 using top-k selection; (3) **D^{2}Pruner** (Zhang et al., 2025): state-of-the-art method using offline bias prior (computed from 1000 COCO images) with MIS selection at layer 2.

Table 1: Main results on RefCOCO visual grounding benchmarks (10% token retention). Our method (L12 + MIS) achieves 66.32% average accuracy, surpassing D^{2}Pruner by +11.29 points without offline calibration. Best pruning results in **bold**. †Requires offline calibration on 1000 COCO images.

Method	RefCOCO			RefCOCO+			RefCOCOg		Avg
	val	testA	testB	val	testA	testB	val	test	
No Pruning (100%)	74.34	77.13	71.21	70.35	75.13	65.08	70.55	72.71	72.06
FastV (L2)	35.63	40.09	32.05	30.80	34.19	27.74	34.44	35.83	33.85
D^{2}Pruner (L2)†	57.04	61.76	54.13	50.57	56.72	47.86	55.70	56.49	55.03
Ours (L12)	**69.24**	**72.69**	**64.18**	**64.21**	**69.82**	**57.41**	**65.36**	**67.63**	**66.32**

Table 2: Ablation study: Effect of ratio-based debiasing at different pruning layers. Ratio normalization uniformly degrades performance compared to raw attention. All methods use MIS selection with 10% retention.

Configuration	RefCOCO	RefCOCO+	RefCOCOg	Overall	Δ vs Raw
Raw MIS (L4)	41.10	35.87	39.35	**38.70**	—
Ratio (L4, K_s=3)	14.67	11.96	12.99	13.24	-25.46
Raw MIS (L12)	68.70	63.82	66.50	**66.32**	—
Ratio (L12, K_s=3)	56.62	52.87	54.29	54.63	-11.69
Ratio (L12, K_s=2)	63.00	59.15	61.16	61.10	-5.22
Weighted Combo (L12)	66.44	61.98	64.32	64.24	-2.08

Implementation. All methods use 10% token retention (keeping 26 of 256 visual tokens). For our method and D^{2}Pruner, we use identical MIS hyperparameters: $r_{\text{pivot}} = 0.7$, $\theta_{\text{sim}} = 0.8$, $\alpha = 0.5$. Experiments were conducted on 4×A100-80GB GPUs with deterministic decoding.

4.2 Main Results

Table 1 presents the main results. Our deep-layer attention pruning achieves 66.32% average grounding accuracy, surpassing D^{2}Pruner (55.03%) by +11.29 points while eliminating the need for offline calibration. The improvement is consistent across all 8 benchmark splits, with gains ranging from +9.55 points (RefCOCO+ testB) to +13.64 points (RefCOCO+ val).

Compared to the no-pruning upper bound (72.06%), our method retains 92.0% of the original performance while reducing visual tokens by 90%. In contrast, D^{2}Pruner retains only 76.4% of no-pruning performance, and FastV retains just 47.0%. This demonstrates that deep-layer attention provides substantially better token importance estimates than shallow-layer attention, even with sophisticated debiasing.

4.3 Ablation Study: Why Ratio Debiasing Fails

Table 2 examines why ratio-based debiasing fails. The ratio formula $A_{\text{mid}}/(A_{\text{shallow}} + \epsilon)$ degrades performance at every tested configuration. At layer 4, ratio debiasing causes catastrophic failure: 13.24% accuracy compared to 38.70% with raw attention (-25.46 points). Even at layer 12 where attention is more semantically meaningful, ratio debiasing still hurts: -11.69 points with $K_s = 3$ and -5.22 points with $K_s = 2$. A softer weighted combination approach also underperforms raw attention by -2.08 points.

These results demonstrate that the ratio formula does not remove positional bias—it destroys useful saliency signal. The consistent degradation across all configurations refutes the hypothesis that shallow-layer attention can serve as a debiasing prior.

Table 3: Phase-0 diagnostic metrics across layers. Position correlation is uniformly weak (∼0.17) at all layers, well below the 0.3 threshold for meaningful positional bias. This explains why ratio-based debiasing fails.

Layer	Prompt Stability	Position Correlation	Entropy
1	0.997	0.143	5.90
2	0.995	0.155	5.55
3	0.997	0.175	4.71
4	0.997	0.176	4.85
8	0.979	0.179	5.81
12	0.928	0.175	5.04

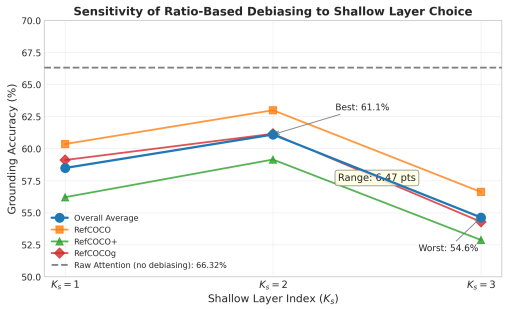

Figure 2: Sensitivity of ratio-based debiasing to shallow layer choice (K_s). Performance varies by 6.47 points across K_s values. Raw attention (dashed line) consistently outperforms all ratio variants.

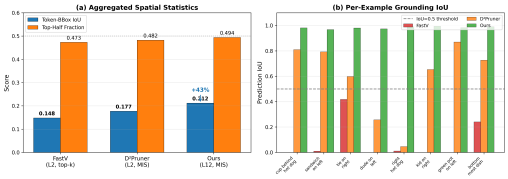

Figure 3: Spatial analysis of token retention. Our method achieves 43% higher token-bbox IoU (0.212) than FastV (0.148), indicating better concentration of retained tokens around referred objects.

4.4 Diagnostic Analysis: No Positional Bias to Correct

Table 3 presents diagnostic metrics from our Phase-0 analysis on 30 COCO images with 5 prompt templates (150 forward passes). The key finding is that position correlation (Spearman ρ between attention and token position) is uniformly weak across all layers, ranging from 0.143 to 0.179. All values are well below the 0.3 threshold typically considered indicative of meaningful positional bias.

This explains why ratio-based debiasing fails: there is no position-specific bias to correct in InternVL2.5-8B's attention patterns. The shallow-layer attention is highly prompt-stable (cosine similarity > 0.99), but this stability reflects consistent image-level saliency rather than positional artifacts. Dividing by this stable signal removes useful information rather than bias.

4.5 Qualitative Analysis

Figure 2 shows that ratio-based debiasing is sensitive to the choice of shallow layer K_s, with performance varying by 6.47 points across configurations. Importantly, raw attention without any debiasing (dashed line at 66.32%) consistently outperforms all ratio variants, confirming that debiasing is unnecessary when using deep-layer attention.

Figure 3 presents spatial analysis of token retention across 8 RefCOCO examples. Our method achieves 43% higher token-bbox IoU (0.212) compared to FastV (0.148) and 20% higher than D^2Pruner (0.177). This indicates that deep-layer attention produces more semantically focused token selection, with retained tokens better concentrated around the referred objects.

5 Conclusion

We presented a simple yet effective approach to visual token pruning in vision-language models: using attention from deeper layers (L12) rather than shallow layers (L2). This eliminates the

need for offline calibration data required by prior methods like D^2Pruner, while achieving superior performance—66.32% grounding accuracy on RefCOCO benchmarks, surpassing D^2Pruner by +11.29 points. Our diagnostic analysis reveals that shallow-layer attention in InternVL2.5-8B lacks the positional bias assumed by debiasing approaches, explaining why ratio-based normalization degrades rather than improves performance. The key insight is that the solution to attention-based pruning is simpler than previously thought: deeper layers produce semantically richer attention patterns that do not require debiasing.

Limitations. Our evaluation is limited to one model family (InternVL) and one task type (visual grounding). Future work should validate these findings across diverse VLM architectures and tasks including VQA and document understanding.

REFERENCES

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. Token merging: Your vit but faster. *ArXiv*, abs/2210.09461, 2022.

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. pp. 19–35, 2024a.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Zhong Muyan, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. Intern vl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24185–24198, 2023.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Hui Deng, Jiaye Ge, Kaiming Chen, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahu Lin, Yunfeng Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *ArXiv*, abs/2412.05271, 2024b.

Tri Dao, Daniel Y. Fu, Stefano Ermon, A. Rudra, and Christopher R'e. Flashattention: Fast and memory-efficient exact attention with io-awareness. *ArXiv*, abs/2205.14135, 2022.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, M. Minderer, G. Heigold, S. Gelly, Jakob Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

Mark Endo, Xiaohan Wang, and S. Yeung-Levy. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. *ArXiv*, abs/2412.13180, 2024.

Yizhang Jin, Jian Li, Yexin Liu, Tianjun Gu, Kai Wu, Zhengkai Jiang, Muyang He, Bo Zhao, Xin Tan, Zhenye Gan, Yabiao Wang, Chengjie Wang, and Lizhuang Ma. Efficient multimodal large language models: a survey. *Visual Intelligence*, 3, 2024.

Kaiyuan Li, Xiaoyue Chen, Chen Gao, Yong Li, and Xinlei Chen. Balanced token pruning: Accelerating vision language models beyond local optimization. *ArXiv*, abs/2505.22038, 2025.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *ArXiv*, abs/2304.08485, 2023.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. *ArXiv*, abs/2403.15388, 2024.

Gaurav Shinde, Anuradha Ravi, Emon Dey, Shadman Sakib, Milind Rampure, and Nirmalya Roy. A survey on efficient vision-language models. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 15, 2025.

Zhichao Sun, Yidong Ma, Gang Liu, Yibo Chen, Xu Tang, Yao Hu, and Yongchao Xu. Ivc-prune: Revealing the implicit visual coordinates in lvlms for vision token pruning. 2026.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. Efficient streaming language models with attention sinks. *ArXiv*, abs/2309.17453, 2023.

Long Xing, Qidong Huang, Xiao wen Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *ArXiv*, abs/2410.17247, 2024.

Hao Yin, Guangzong Si, and Zilei Wang. Lifting the veil on visual information flow in mllms: Unlocking pathways to faster inference. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 9382–9391, 2025.

Licheng Yu, Patrick Poirson, Shan Yang, A. Berg, and Tamara L. Berg. Modeling context in referring expressions. *ArXiv*, abs/1608.00272, 2016.

Evelyn Zhang, Fufu Yu, Aoqi Wu, Zichen Wen, Ke Yan, Shouhong Ding, Biqing Qi, and Linfeng Zhang. D2pruner: Debiased importance and structural diversity for mllm token pruning. *ArXiv*, abs/2512.19443, 2025.

Yuan Zhang, Chunkai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *ArXiv*, abs/2410.04417, 2024.

Kai Zhao, Wubang Yuan, Yuchen Lin, Liting Ruan, Xiaofeng Lu, Deng-Ping Fan, Ming-Ming Cheng, and Dan Zeng. Attention debiasing for token pruning in vision language models. 2025.