AdaV: Adaptive Text-visual Redirection for Vision-Language Models

Anonymous ACL submission

Abstract

The success of Vision-Language Models (VLMs) often relies on high-resolution schemes that preserve image details, while these approaches also generate an excess of visual tokens, leading to a substantial decrease in model efficiency. A typical VLM includes a visual encoder, a text encoder, and an LLM. Recent studies suggest pruning visual tokens based on visual and textual priors to accelerate VLMs without additional training costs. However, these methods often overlook prompt semantics or suffer from biased self-attention in the LLM. Inspired by the efficient mechanisms of the human brain for multimodal understanding, we introduce AdaV, a novel training-free visual token pruning method. By emulating the neural pathways that preprocess visual and auditory information before the reasoning stage, we shift text-guided visual attention redirection to the pre-LLM stage, which reduces biased token pruning and enhances model robustness with a limited visual token budget. A Self-adaptive Cross-modality Attention Redirection (SCAR) module is further proposed that effectively merges and redirects visual attention with text-to-image attention. Extensive experiments on seven challenging benchmarks demonstrate that our AdaV achieves SOTA performance in training-free VLM acceleration and can be plug-and-play on various VLMs. We plan to open-source the code upon publication.

1 Introduction

In recent years, vision-language models (VLMs) have demonstrated exceptional performance in various visual-grounded tasks. Despite their impressive achievements, the computational cost associated with VLMs remains a significant challenge for practical deployment. A key factor contributing to this cost is the large number of visual tokens required. For instance, LLaVA-NEXT models (Liu et al., 2024) utilize 2,880 visual tokens for single-image tasks, which may significantly exceed the number of tokens typically used in text prompts.

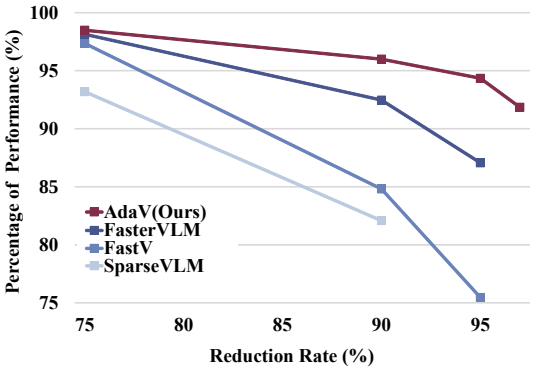

Figure 1: The comparison of training-free VLM acceleration methods on LLaVA-NEXT-7B shows that AdaV achieves state-of-the-art (SOTA) performance.

Many research efforts have focused on pruning redundant visual tokens to accelerate VLMs without additional training. FastV (Chen et al., 2024a) observes that the distribution of attention weights among visual tokens tends to cluster, allowing for the ranking and retention of only the top-ranked tokens in the LLM layers. SparseVLM (Zhang et al., 2024b) selects the keywords from the text and reserves key visual tokens within the self-attention layers of the LLM. However, FasterVLM (Zhang et al., 2024a) highlights that these methods suffer from biased text-to-image attention of the LLM and may not accurately reflect the importance of visual tokens. To address this, FasterVLM proposes utilizing class attention extracted from the visual encoder as a significance metric for visual token pruning. However, it fails to recall non-salient yet semantically relevant visual information.

Previous research has demonstrated that insights from the mechanisms of the human brain can inspire advancements in intelligent systems (Rivest et al., 2004; Hassabis et al., 2017). The human brain tackles multimodal understanding through a

series of steps: (I) processing visual and linguistic information separately within their respective cortexes, (II) matching information and redirecting attention, primarily occurring in the temporoparietal junction (TPJ), and (III) engaging in higher-order thinking and response generation within the prefrontal cortex (PFC) (Miller and Cohen, 2001; Grill-Spector and Weiner, 2014; Doricchi et al., 2022a). The TPJ, situated at the convergence of the temporal and parietal lobes, is crucial for various cognitive functions, including the reorientation of attention and the matching of visual and auditory language inputs. The encoded visual and linguistic information undergoes initial cross-modal attention reorientation in regions such as the TPJ. This stage of processing is distinct from the subsequent activities that occur in the PFC. After the TPJ's involvement, the PFC engages in higher-order cognitive processes, including decision-making and judgment. This workflow enables the brain to concentrate on essential visual information guided by linguistic cues (Lupyan et al., 2020; Doricchi et al., 2022b).

Inspired by these cognitive processes, we propose AdaV, a novel training-free acceleration method that emulates the mechanisms of the human brain. As depicted in Fig. 2, we decompose the VLM into four components corresponding to specific brain regions: (1) the visual encoder (red) mirrors the function of the visual cortex, (2) the text encoder (purple) aligns with the temporal lobe, responsible for comprehending language semantics, (3) the LLM (green) parallels the prefrontal cortex (PFC), which is involved in cognitive processing and responses, and (4) the Self-adaptive Cross-modality Attention Redirection (SCAR) module (blue) in the pre-LLM stage mimics the TPJ's function to integrate multimodal information. First, we extract visual attention from the self-attention layers within the visual encoder. Next, embedded text prompts query the visual embeddings to obtain text-to-image attention. We then measure the overall significance of potential visual token collections using a geometric average of both visual and text-to-image attention and employ a one-step optimization process to determine the optimal visual token collection. Our validation experiments demonstrate that, compared with the text-to-image attention extracted from the self-attention layers of the LLM, the attention in the pre-LLM stage mitigates the attention bias, and effectively reflects the significance of visual tokens. Additionally, extensive experiments show that our AdaV achieves state-of-the-art (SOTA) performance on multiple benchmarks and is even comparable to fine-tuning methods such as VisionZip (Yang et al., 2024). Our contributions are summarized as follows:

I. We propose AdaV, a novel training-free VLM acceleration method that effectively speeds up VLMs while maintaining the model's effectiveness.

II. Inspired by the human brain's multimodal information-processing pathways, we introduce visual attention redirection to the pre-LLM stage and reveal the necessity and feasibility of this design. Experiments demonstrate it significantly enhances the model's performance given a limited visual token budget.

III. We propose a Self-adaptive Cross-modality Attention Redirection (SCAR) module that effectively redirects visual attention via text-to-image attention for effective visual token pruning.

IV. Extensive experiments on seven benchmarks show that AdaV achieves SOTA performance and is plug-and-play on various VLMs.

2 Related work

2.1 Vision language models (VLMs)

Significant progress has been made in the development of VLMs. LLaVA (Yifan et al., 2023) is the first approach to effectively combine large language models (LLMs) with foundational vision models. The initial models in the LLaVA family only utilize a single image as input, resulting in 576 visual tokens for an image. However, this approach often led to significant information loss, thereby limiting model performance. To retain detailed vision information of the input, the subsequent VLMs propose dynamic resolution schema to the input image, enhancing multi-modal capabilities (Lin et al., 2023; Wang et al., 2024b; Chen et al., 2024b).

2.2 VLM acceleration with token pruning

Token pruning is a straightforward solution for accelerating transformer models and is widely used in different deep learning tasks (Kim et al., 2022; Nawrot et al., 2023; Zhong et al., 2023; Wang et al., 2024a). Recent works have adopted this concept to accelerate VLMs. Chen et al. (2024a); Ye et al. (2024) propose measuring the significance of visual tokens based on self-attention extracted from layers within LLMs. FasterVLM suggests that text-to-image attentions in LLM layers are biased, and shift to the ends of input image tokens, leading to

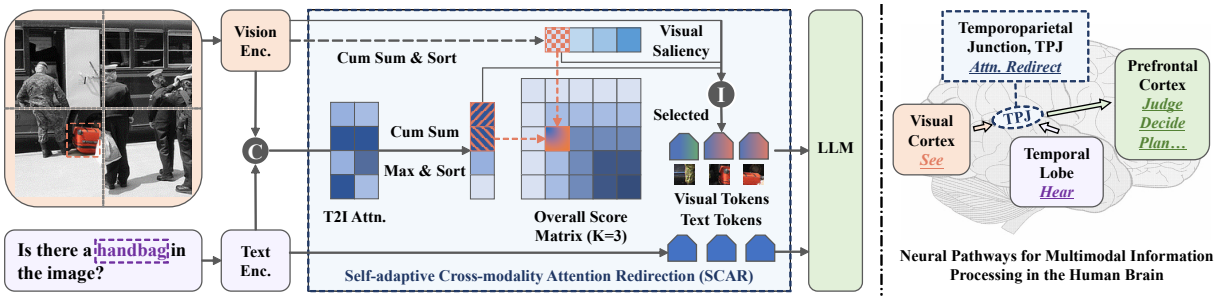

Figure 2: The overall framework of the AdaV. Inspired by human brain information-processing pathways, we redirect the visual attention by the SCAR module before the LLM stage for robust and effective visual token pruning.

inaccuracies. Instead, they propose using visual attention to prune non-salient visual tokens. Some approaches also fine-tune the VLMs for effectiveness. For example, VisionZip (Yang et al., 2024) proposes to finetune the MLP projector of the vision encoder for the reserved visual tokens.

3 Method

3.1 Overall framework

As depicted in Fig. 2, our overall framework decomposes the VLMs into four parts: the visual encoder, the text encoder, the LLM, and the SCAR module prior to the LLM. Input images and text prompts are encoded separately via the corresponding encoder. The SCAR module then redirects the visual attention via text-to-image attention for visual token pruning. The retained visual tokens, along with the text tokens, are then forwarded to the LLM for generating responses.

3.2 Necessity and feasibility of visual attention redirection in the pre-LLM stage

Recent studies have demonstrated that text-to-image attention within the LLM is biased, leading to sub-optimal visual token filtering. These studies propose introducing text-agnostic visual attention for token pruning, which significantly boosts model performance (Zhang et al., 2024a; Yang et al., 2024). Consequently, before formally introducing the proposed AdaV, we address the following question:

Is visual attention alone sufficient for visual token pruning?

We conducted experiments on five benchmarks to answer this question: MME, MM-Vet, TextVQA, POPE, and GQA. We employed the SAM-2 model (Ravi et al., 2024) to segment objects based on text prompts and used the CLIP vision encoder (Radford et al., 2021) to generate visual attention for

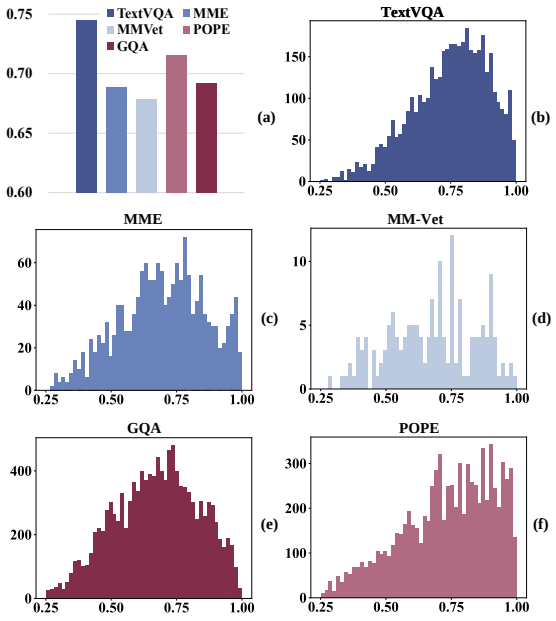

Figure 3: The average AUC on different datasets (a) and the distribution of AUC on each dataset (b∼f). Despite the AUC on all datasets being large, there exist samples that visual attention fails to handle.

input images. The area under the curve (AUC) assesses the consistency between visual attention and text-grounded segmentation. Detailed experimental procedures are provided in the Appendix (A.2). As illustrated in Fig. 3 (a), the average AUC across datasets exceeds 0.65, indicating that these tasks are generally grounded in salient visual cues. However, as shown in Fig. 3 (b∼f), despite visual attention being a strong priority for informative visual token selection, some tasks exhibit AUC ≤ 0.5, where visual attention fails to perform effectively. Thus, we conclude:

No. Although visual attention is impactful, there are tasks it fails to handle.

Since visual attention alone is not sufficient, text-guided attention redirection is needed to focus on

non-salient yet semantically important visual information. However, FasterVLM (Zhang et al., 2024a) validates that the text-to-image attention extracted from the LLM stage is biased, shifting to the ends of the image and thus failing to identify the visual tokens related to the text prompts accurately. The widely utilized CLIP model encodes the vision and text input separately and effectively aligns their embeddings, which converge to the human brain information processing flow before PFC. However, unlike CLIP, which utilizes an entire transformer model to encode text, the VLM's text encoder (text embedding layer) is much smaller. Therefore, prior to adopting text-guided attention redirection to the pre-LLM stage, we need to address the following question:

Is text embedding layer sufficient for generating well-aligned representations to visual embeddings?

We address this question in two steps. First, we measure the distribution of text-to-image similarity for embedded text tokens and visual embeddings using a subset of the LLaVA dataset (Liu et al., 2023), following FasterVLM. We employed two metrics to assess alignment: normalized cosine similarity and normalized inner product. The results, visualized in Fig. 4, show no attention shift in the pre-LLM stage. However, the normalized inner product produces significant outliers, potentially degrading model performance. In contrast, normalized cosine similarity demonstrates a more uniform distribution, effectively mitigating outliers and offering greater reliability.

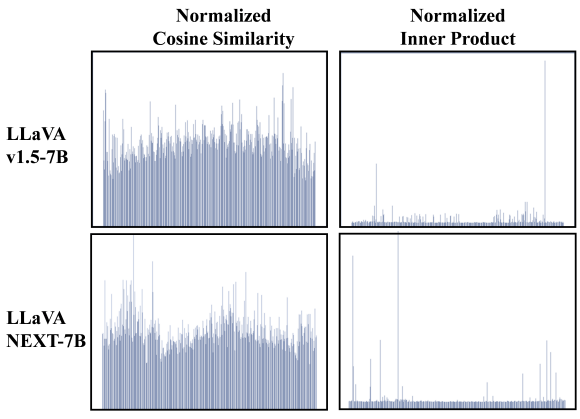

Figure 4: Text-to-image similarity distribution of LLaVA-v1.5-7B and LLaVA-NEXT-7B.

Having established that text-to-image similarity in pre-LLM layers does not exhibit attention shifts, we next investigate whether text embeddings are sufficiently aligned with visual embeddings to facilitate effective visual token selection. To achieve this, we use the least number of reserved tokens to cover one visual token relevant to the question as our validation metric. Specifically, if the N_{th} visual token is the first visual token relevant to the question, N serves as the least number of reserved tokens. Following the methodology of FasterVLM, we conducted experiments on a subset of the LLaVA data collection. We utilize the same pipeline as described in Sec. 3.2 to determine the relevant visual tokens. Our findings, shown in Fig. 5, indicate that text-to-image similarity requires fewer reserved visual tokens to cover at least one relevant visual token, compared to visual attention. We could conclude as follows:

Yes. The text embedding layer can generate text representations that are aligned with the visual embeddings while mitigating attention bias.

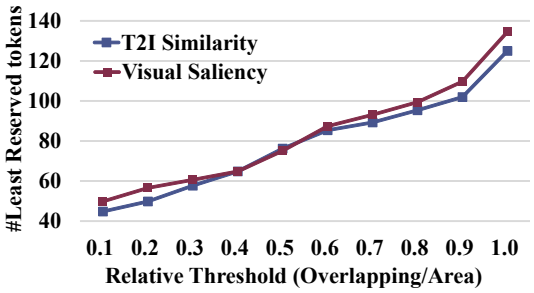

Figure 5: The average of the least number of reserved visual tokens to select at least one prompt-related visual token, validated on LLaVA-1.5-7B.

3.3 Self-adaptive Cross-modality Attention Redirection (SCAR)

Redirecting visual attention changes the selection of visual tokens from relying solely on text-agnostic visual attention to a co-dependent approach that incorporates both visual attention and text-to-image attention. As noted in ViT (Dosovitskiy et al., 2020), the [CLS] token encapsulates global information. We thus identify the visual attention with the self-attention weight between the [CLS] token and image patches, which is called "image attention". Note that image attention distinct from the concept of visual attention. Formally, let the input text embeddings be $T_E \in \mathbb{R}^{N_T \times D}$, visual embeddings be $T_V \in \mathbb{R}^{N_{\text{img}} \times N_I \times D}$, and the [CLS] token be $C \in \mathbb{R}^{N_{\text{img}} \times D}$, where N_T, N_{img}, and N_I represent the number of text tokens, images, and visual tokens per image, respectively. Denote the image attention of the i_{th} image as S_i^C, which

could be calculated as follows:

$$S_i^C = \text{Softmax}\Big(\frac{C_i W_Q \big((T_V)_i W_K\big)^T}{\lambda}\Big). \quad (1)$$

A straightforward approach for redirecting visual attention involves selecting visual tokens based on both image attention and the similarity between text and image tokens, denoted as S_i^{T2I}, simultaneously. This similarity could be formulated as follows:

$$S_{i,j}^{\text{T2I}} = \max_k \Big(\frac{(T_E)_k (T_V)_{i,j}^T}{||(T_E)_k||_2 ||((T_V)_{i,j})||_2}\Big). \quad (2)$$

The text-agnostic visual attention is redirected from visual tokens selected solely by image attention to those chosen as follows:

$$\begin{cases} I_i = \Big[\text{argtop-K}\big(S_{i,:}^C\big); \text{argtop-K}\big(S_{i,:}^{\text{T2I}}\big)\Big] \\ (\hat{T}_V)_i = (T_V)_{I_i} \end{cases}, \quad (3)$$

in which $(\hat{T}_V)_i$ represents the retained tokens for the i_{th} image.

Since the effectiveness of the text-agnostic visual attention on identifying informative visual tokens could vary across samples and tasks, redirection formulated in Eq. 3 could be sub-optimal. SCAR optimizes the mixed significance of the valid collections of visual tokens to determine attention redirection adaptively. Since the cosine similarity of text and visual embeddings has a different distribution to the image attention which is extracted from self-attention layers, we first re-weight the similarity as follows:

$$\tilde{S}_{i,j}^{\text{T2I}} = \max_k \left(\text{Softmax}\Big(\frac{(T_E)_k (T_V)_{i,j}^T}{||(T_E)_k||_2 ||((T_V)_{i,j})||_2 \tau}\Big)\right), \quad (4)$$

in which τ is a hyper-parameter, which is set to 0.01. The re-weighted text-to-image similarity is called "T2I attention". For simplicity, we merge the dimensionality of images and tokens per image, so S^C and $\tilde{S}^{\text{T2I}}$ are reshaped to $(N_{\text{img}} \times N_I,)$. Maximizing the T2I and image attention of the selected visual tokens is equivalent to maximizing the following objective:

$$\sum_{m \in \mathcal{M}} \tilde{S}_m^{\text{T2I}} + \sum_{n \in \mathcal{N}} S_n^C, \quad s.t. \quad |\mathcal{M}| + |\mathcal{N}| = K, \quad (5)$$

in which $\mathcal{M}$ and $\mathcal{N}$ represent the sets of selected visual token indices based on T2I attention and image attention, respectively. Since the distributions of T2I attention and image attention are different, and only a small group of tokens are retained, maximizing the objective in Eq. 5 may result in solely depending on an individual metric, which is not expected. To address this, we utilize the geometric mean of the metrics to measure the importance of the selected tokens:

$$\sqrt{\sum_{m \in \mathcal{M}} \tilde{S}_m^{\text{T2I}} \sum_{n \in \mathcal{N}} S_n^C}, \quad s.t. \quad |\mathcal{M}| + |\mathcal{N}| = K, \quad (6)$$

We start by sorting $\tilde{S}^{\text{T2I}}$ and S^C, resulting in the sorted scores $\hat{S}^{\text{T2I}}$ and $\hat{S}^C$. Next, we calculate the cumulative summations of these sorted scores, denoted as $\mathbf{a}$ and $\mathbf{b}$ respectively, as follows:

$$\begin{aligned} &\mathbf{a}_0 = 0, \mathbf{b}_0 = 0, \\ &\mathbf{a}_t = \sum_{m=0}^{t-1} \hat{S}_m^{\text{T2I}}, \mathbf{b}_t = \sum_{n=0}^{t-1} \hat{S}_n^C. \end{aligned} \quad (7)$$

Then we calculate the overall metrics as follows:

$$O = \mathbf{a}\mathbf{b}^T. \quad (8)$$

In order for the invalid indices not to be chosen, we utilize a mask M to set the elements of O corresponding to such indices to zero. Specifically, the mask M could be calculated as follows:

$$M_{m,n} = \begin{cases} 1, m+n \leq K \\ 0, otherwise \end{cases}. \quad (9)$$

Then the number of tokens selected by T2I attention and image attention could be determined as follows:

$$U, V = \underset{m,n}{\text{argmax}}\{(O \otimes M)_{m,n}\}. \quad (10)$$

Finally, the SCAR module redirects the original text-agnostic visual attention to the following visual tokens:

$$\begin{cases} \mathcal{M} = \{m | \text{rank}(\tilde{S}_m^{\text{T2I}}) \leq U\} \\ \mathcal{N} = \{n | \text{rank}(S_n^C) \leq V\} \\ \hat{T}_V = \{(T_V)_k\}_{k \in \mathcal{M} \cup \mathcal{N}} \end{cases}, \quad (11)$$

in which $\text{rank}(A_i)$ returns the position of the element A_i after sorting A in a descending order. We then sort the preserved tokens according to their original position.

4 Experiments

4.1 Implementation details

We evaluate the proposed approach on the LLaVA-v1.5-7B and LLaVA-NEXT models (7B, 13B, and 34B parameters) across seven distinct VLM benchmarks: GQA (Hudson and Manning, 2019), SQA (Lu et al., 2022), MME (Fu et al., 2024), MMBench (Liu et al., 2025), MM-Vet (Yu et al., 2023), TextVQA (Singh et al., 2019), and Pope (Yifan et al., 2023). All experiments were conducted using the NVIDIA A100-80G GPU.

Table 1: Comparison with SOTA approaches on LLaVA-NEXT-7B. † means that we report both the perception-only score and the summation of the perception score and the cognition score in parenthesis. ‡ with a gray background means the model is fine-tuned, which is expected to be stronger. "Average" represents the overall performance.

Method	Average	GQA	SQA-IMG	TextVQA	POPE	MME	MMB	MM-Vet
LLaVA-NEXT-7B	100.00%	62.93	69.66	59.59	86.32	1513.78 (1842.00)	67.70	42.60
Reduction Rate ≈ 75%								
FastV	97.35%	60.38	69.81	58.39	83.09	1477.31	65.64	41.10
SparseVLM	93.19%	60.88	67.48	58.08	70.99	1446.10	63.83	38.00
FaseterVLM	98.14%	61.31	68.82	59.33	85.50	1480.68	67.35	40.40
AdaV (Ours)	**98.49%**	**62.04**	**69.31**	**58.37**	**87.20**	**1509.36**	**67.35**	**39.70**
VisionZip	97.75%	61.30	68.10	60.20	86.30	1702.00	66.30	
AdaV (Ours)	**99.13%**	**62.04**	**69.31**	**58.37**	**87.20**	**1810.07**	**67.35**	
VisionZip+FT‡	99.00%	62.40	67.90	60.80	87.60	1778.00	65.90	
Reduction Rate ≈ 90%								
FastV	84.81%	55.86	69.26	55.69	71.66	1282.86	61.60	22.70
SparseVLM	82.08%	56.12	68.62	51.97	63.23	1332.22	54.47	24.70
FaseterVLM	92.47%	58.12	68.12	57.57	80.00	1370.11	63.32	35.70
AdaV (Ours)	**96.00%**	**60.65**	**68.57**	**57.09**	**85.98**	**1503.25**	**66.32**	**36.00**
VisionZip	95.07%	59.30	67.30	58.90	82.10	1702.00	63.10	
AdaV (Ours)	**97.77%**	**60.65**	**68.57**	**57.09**	**85.98**	**1812.89**	**66.32**	
VisionZip+FT‡	97.40%	61.00	67.50	59.30	86.20	1770.00	64.40	
Reduction Rate ≈ 95%								
FastV	75.46%	49.83	68.52	51.85	51.66	1079.46	54.90	21.90
FaseterVLM	87.06%	54.73	68.86	55.97	72.89	1225.96	60.48	31.90
AdaV (Ours)	**94.35%**	**58.53**	**68.91**	**55.11**	**85.25**	**1452.91**	**65.20**	**36.20**
VisionZip	90.75%	55.50	68.30	56.20	74.80	1630.00	60.10	
AdaV (Ours)	**95.62%**	**58.53**	**68.91**	**55.11**	**85.25**	**1736.12**	**65.20**	
VisionZip+FT‡	94.80%	58.20	67.50	57.30	83.40	1699.00	63.90	

4.2 Comparison with SOTA approaches

We compare our proposed approach with other state-of-the-art (SOTA), training-free token pruning methods. Due to variations in benchmark datasets, reduction rates, and evaluation metrics across different studies (e.g., VisionZip uses the sum of perception and cognition scores, while FasterVLM focuses solely on perception scores), we present our detailed comparisons in Table 1 for clarity, specifically for the LLaVA-NEXT-7B model. Additionally, Table 2 briefly demonstrates the effectiveness of the proposed AdaV on other VLMs, with detailed comparisons available in the Appendix (A.5). Our approach achieves state-of-the-art performance among training-free methods and even surpasses the fine-tuned VisionZip. It shows remarkable robustness, particularly when preserving less than 10% of visual tokens.

Table 2: Comparison with SOTA approaches

Method	Reduction Rate 75%	90%	95%
LLaVA-1.5-7B			
FastV	94.67%	86.26%	72.48%
SparseVLM	93.22%	78.87%	65.85%
FaseterVLM	98.32%	92.91%	87.76%
AdaV (Ours)	**97.83%**	**93.59%**	**88.32%**
LLaVA-NEXT-13B			
FaseterVLM	97.57%	92.79%	86.52%
AdaV (Ours)	**97.75%**	**95.40%**	**93.14%**
LLaVA-NEXT-34B			
FaseterVLM	/	89.29%	83.90%
AdaV (Ours)	/	**91.85%**	**88.11%**

4.3 Ablation study

Overall ablation We conduct an overall ablation study of the proposed approach. As demonstrated in Table 3, the T2I attention significantly boosts the model performance, especially when the number of retained tokens is small. Additionally, the proposed SCAR module further improves the model's performance by over 1.0% at reduction rates exceeding 90%.

Table 3: Ablation study of main modules on LLaVA-NEXT-7B

Model	Reduction Rate (%) 75	90	95
AdaV (Ours)	**98.49%**	**96.00%**	**94.35%**
-SCAR	98.40%	94.89%	92.62%
-T2I Attention	98.18%	92.47%	87.06%

Detailed ablation results on specific datasets To further understand the influence of the proposed mechanisms, we validated the model on two datasets: POPE and MMBench. The results are presented in Tables 4 and 5. By combining image attention with T2I attention, the model effectively redirected text-agnostic visual attention to question-related visual information, thereby enhancing performance. However, this simple redirection occasionally led to performance degradation, indicating that the selection might be sub-optimal. The proposed SCAR module offers an effective integration of image and T2I attention, significantly improving upon the simple redirection method.

Table 4: Ablation study on the Pope dataset. "SCAR", "T2I" and "IA" demonstrate the SCAR module, T2I attention and image attention, respectively.

IA	T2I	SCAR	Reduction Rate (%) 75	90	95
✓	×	×	85.50	80.00	72.89
✓	✓	×	87.07	85.52	84.04
✓	✓	✓	**87.20**	**85.98**	**85.25**

Attention dependency analysis We further analyzed the attention dependency across different datasets, with results illustrated in Fig. 6. Among the figure, if a curve is positioned on the left side, the model relies more on T2I attention; otherwise, it depends more on image attention. Our analysis reveals that the model tends to rely more on image attention for tasks requiring optical character recognition, such as TextVQA and MM-Vet. Conversely, for tasks primarily involving natural images, the SCAR module redirects more visual attention to information relevant to the linguistic input. This demonstrates that the proposed SCAR module effectively determines the balance between image and T2I attention, enhancing the performance of the VLMs upon visual token pruning.

Table 5: Ablation study on the MMBench dataset. "SCAR", "T2I" and "IA" demonstrate the SCAR module, T2I attention and image attention, respectively.

IA	T2I	SCAR	Reduction Rate (%) 75	90	95
✓	×	×	67.35	63.32	60.48
✓	✓	×	66.32	65.80	64.17
✓	✓	✓	**67.35**	**66.32**	**65.20**

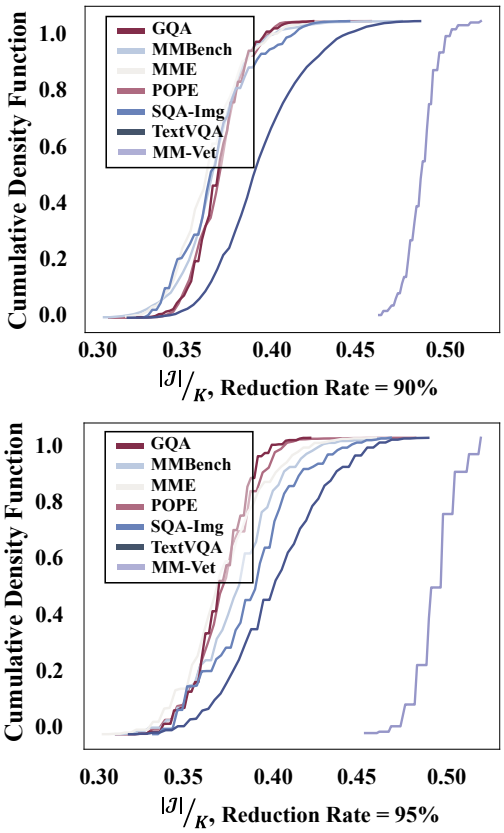

Figure 6: The cumulative density function (CDF) of the proportion of image attention-oriented tokens ($\mathcal{J}$) on different benchmarks, validated on LLaVA-NEXT-7B.

Figure 7: Visualization of selected tokens. Transparent patches indicate unselected tokens. Comparing columns 2 and 4 shows that AdaV successfully identifies non-salient yet relevant visual tokens, which FasterVLM fails to accomplish. The comparison between columns 4 and 5 demonstrates AdaV's ability to redirect attention based on the text prompt, which FasterVLM fails to achieve.

4.4 Visualization of selected tokens

We further visualize the selected tokens of the FasterVLM and the proposed approach in Fig. 7. Since the FasterVLM approach is text-agnostic, the selected visual tokens are consistent with a certain input image, which results in the VLM only accessing the salient objects, and failing to allocate the cases in which the user prompts are about non-salient objects in the image. On the contrary, the proposed approach effectively leverages the strength of both image attention and T2I attention. As depicted in Fig. 7, the proposed approach could draw attention to the non-salient visual information, according to the guidance of the text prompts. Furthermore, FasterVLM focuses on exactly the same visual information for a certain input, discarding the change of the question. On the contrary, the proposed AdaV is capable of shifting its attention according to the text prompt.

5 Conclusion

In this study, we introduce AdaV, a training-free approach designed to accelerate VLMs by emulating the multimodal information processing pathways of the human brain. Our method positions text-guided visual attention redirection before the LLM, effectively mitigating biased and text-agnostic token preservation. Additionally, we present the Self-adaptive Cross-modality Attention Redirection (SCAR) module, which adaptively integrates and redirects visual attention in conjunction with text-to-image attention. Extensive experiments demonstrate that AdaV achieves state-of-the-art performance compared to existing approaches for training-free VLM acceleration and is plug-and-play on various VLMs.

6 Limitations

In this section, we discuss the limitations of the proposed approach. Although AdaV effectively demonstrates the benefits of visual token pruning, it relies heavily on the alignment between text embeddings and visual information. Our visualizations indicate that many preserved visual tokens are still redundant and irrelevant to the text prompt, which constrains the model's performance and efficiency. Further exploration into the nature of the visual encoder and text embeddings is necessary to enhance visual token pruning.

References

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024a. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *Computer Vision – ECCV 2024*, pages 19–35, Cham. Springer Nature Switzerland.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198.

Fabrizio Doricchi, Stefano Lasaponara, Mariella Pazzaglia, and Massimo Silvetti. 2022a. Left and right temporal-parietal junctions (tpjs) as "match/mismatch" hedonic machines: A unifying account of tpj function. *Physics of Life Reviews*, 42:56–92.

Fabrizio Doricchi, Stefano Lasaponara, Mariella Pazzaglia, and Massimo Silvetti. 2022b. Left and right temporal-parietal junctions (tpjs) as "match/mismatch" hedonic machines: A unifying account of tpj function. *Physics of Life Reviews*, 42:56–92.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, Yunsheng Wu, and Rongrong Ji. 2024. Mme: A comprehensive evaluation benchmark for multimodal large language models. *Preprint*, arXiv:2306.13394.

Kalanit Grill-Spector and Kevin S Weiner. 2014. The functional architecture of the ventral temporal cortex and its role in categorization. *Nature Reviews Neuroscience*, 15(8):536–548.

Demis Hassabis, Dharshan Kumaran, Christopher Summerfield, and Matthew Botvinick. 2017. Neuroscience-inspired artificial intelligence. *Neuron*, 95(2):245–258.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Sehoon Kim, Sheng Shen, David Thorsley, Amir Gholami, Woosuk Kwon, Joseph Hassoun, and Kurt Keutzer. 2022. Learned token pruning for transformers. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 784–794.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2023. Vila: On pre-training for visual language models. *Preprint*, arXiv:2312.07533.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. 2025. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems*, 35:2507–2521.

Gary Lupyan, Rasha Abdel Rahman, Lera Boroditsky, and Andy Clark. 2020. Effects of language on visual perception. *Trends in cognitive sciences*, 24(11):930–944.

Earl K Miller and Jonathan D Cohen. 2001. An integrative theory of prefrontal cortex function. *Annual review of neuroscience*, 24(1):167–202.

Piotr Nawrot, Jan Chorowski, Adrian Lancucki, and Edoardo Maria Ponti. 2023. Efficient transformers with dynamic token pooling. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6403–6417.

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

Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, Eric Mintun, Junting Pan, Kalyan Vasudev Alwala, Nicolas Carion, Chao-Yuan Wu, Ross Girshick, Piotr Dollár, and Christoph Feichtenhofer. 2024. Sam 2: Segment anything in images and videos. *arXiv preprint arXiv:2408.00714*.

Françcois Rivest, Yoshua Bengio, and John Kalaska. 2004. Brain inspired reinforcement learning. *Advances in neural information processing systems*, 17.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8317–8326.

Hongjie Wang, Bhishma Dedhia, and Niraj K Jha. 2024a. Zero-tprune: Zero-shot token pruning through leveraging of the attention graph in pretrained transformers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16070–16079.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024b. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *Preprint*, arXiv:2409.12191.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. 2024. Visionzip: Longer is better but not necessary in vision language models. *arXiv preprint arXiv:2412.04467*.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. 2024. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. *arXiv preprint arXiv:2409.10197*.

Li Yifan, Du Yifan, Zhou Kun, Wang Jinpeng, Xin Zhao Wayne, and Ji-Rong Wen. 2023. Evaluating object hallucination in large vision-language models. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, MinQi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. 2024a. [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster. *arXiv preprint arXiv:2412.01818*.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, et al. 2024b. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*.

Qihuang Zhong, Liang Ding, Juhua Liu, Xuebo Liu, Min Zhang, Bo Du, and Dacheng Tao. 2023. Revisiting token dropping strategy in efficient bert pretraining. In *The 61st Annual Meeting Of The Association For Computational Linguistics*.

A Appendix

A.1 Details of involved datasets

All the included datasets are open-sourced and allow academic use. Their details are listed as follows:

GQA benchmark is designed to assess structured comprehension and reasoning skills for scenes presented in images. The questions are crafted based on the scene graphs, ensuring questions are aligned with a specific semantic pathway. Evaluation is conducted on the test-dev set, which comprises 12,578 image-question pairs.

ScienceQA benchmark assesses a model's ability to generalize zero-shot across various scientific domains. Questions of the dataset are systematically categorized by topic, category, and skill level. The vision-grounded subset of the test set is utilized for evaluation, referred to as SQA-IMG, and comprises 2,017 image-question pairs.

TextVQA benchmark focusing on the combination of optical character recognition (OCR) and natural language processing. The images feature a wide range of scenarios, including signs, billboards, and product packaging, all rich in textual content. The validation set that consists of 5,000 image-question pairs is utilized for evaluation.

POPE benchmark is designed to assess hallucinations in large vision-language models by posing questions about the presence of specific objects. For evaluation, an F1 score across three different sampling strategies was applied to the test set, which includes 8,910 image-question pairs.

MME assesses the perceptual and cognitive abilities of multi-modal models through 14 subtasks, including OCR and both coarse- and fine-grained recognition. Performance is measured using the perception and overall scores across 2,374 image-question pairs.

MMBench evaluates multi-modal models across three competence levels. Level one includes perception and reasoning, level two adds six specific capabilities and level three comprises 20 concrete tasks with multiple-choice questions, including 4,377 image-question pairs.

MM-Vet assesses the integration of core vision-language capabilities—recognition, OCR, knowledge, language generation, spatial awareness, and mathematics—across 16 specific tasks, including a total of 218 image-question pairs.

A.2 Pipeline of analyzing the text prompts and the visual salient information

To analyze the alignment of text prompts and the visual salient information, we first segment the text-relevant objects with SAM-2 model. To ensure at least one object is included in each image, we gradually decrease the confidence threshold to 0.2 (with a step of 0.01), until at least one segment mask is obtained. If no mask is obtained, we discard the (question, image) pair. Then we utilize CLIP-ViT-L/14 as the vision encoder to extract the contribution of the visual tokens to the [CLS] token. We do not utilize the ViT for ImageNet classification since the CLIP model has a similar nature to the VLMs. The segment mask is separated into 14×14 non-overlap patches to fit the resolution of the vision encoder. A patch is considered to be related to the text prompt if the mask inside the patch occupies more than 50% of the area of the patch if an extra statement is not made. Then a (confidence, label) pair is created for each visual token to calculate the ROC and AUC. We call this the ROC and AUC of visual attention. Specifically, the confidence is the attention weight, and the label is obtained as follows:

$$\text{label} = \begin{cases} 0, \text{overlap} < 50\% \\ 1, otherwise \end{cases} . \qquad \text{(A1)}$$

A.3 Effect of benchmark pattern on visual token pruning

As illustrated in the paper, the average visual attention AUC of the dataset reflects the pattern of the dataset: whether this dataset tends to contain questions about the visually salient objects of the image. For each dataset, we calculate the relative performance of the FasterVLM and the proposed AdaV, with a 95% reduction rate, on different VLMs. The fitted line demonstrates that there is likely to be a pattern: if the AUC of visual attention is large, depending on it is a better choice, otherwise, the model should depend more on text-to-image attention.

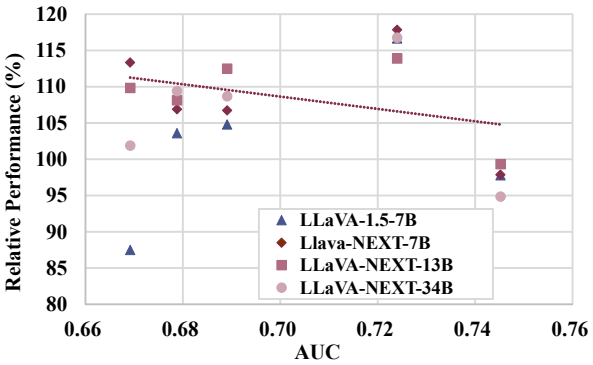

Figure A1: The AUC of visual attention and text-oriented objects versus the relative performance of AdaV and FasterVLM ($\text{Perf}_{\text{AdaV}}/\text{Perf}_{\text{FasterVLM}}$).

A.4 Influence of model scale

We visualize the influence of the model scale for VLM acceleration. As demonstrated in Fig. A2, the increasing scale of the VLMs limits the performance of the visual token pruning in the pre-LLM layers, especially for the text-oriented tasks. As shown in Fig. A3, the degradation caused by the model scale is independent of the visual token pruning method.

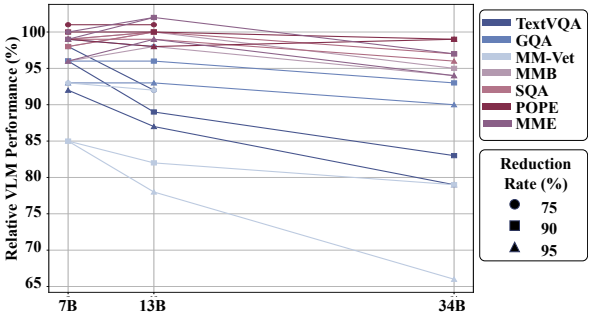

Figure A2: Influence of model scale for visual token pruning across reduction rates and datasets.

A.5 Detailed comparison on LLaVA-v1.5-7B and LLaVA-NEXT-13B/34B

We show a detailed comparison of the token pruning methods on LLaVA-NEXT-13B, LLaVA-

Table A1: Comparison with SOTA approaches on LLaVA-NEXT-13B. † means that we report both the perception-only score and the summation of the perception score and the cognition score in parenthesis. ‡ with a gray background means the model is fine-tuned, which is expected to be stronger.

Method	Average	GQA	SQA-IMG	TextVQA	POPE	MME	MMB	MM-Vet
LLaVA-NEXT-13B	100.00%	65.40	73.60	67.10	86.20	1575.00 (1901.00)	70.00	48.40
Reduction Rate ≈ 75%								
FaseterVLM	97.57%	63.05	72.88	61.67	85.27	1548.06	69.50	48.00
AdaV (Ours)	**97.75%**	**64.26**	**73.33**	**61.93**	**86.70**	**1599.80**	**70.10**	**44.40**
VisionZip	96.93%	63.00	71.20	62.20	85.70	1871.00	68.60	
AdaV (Ours)	**98.82%**	**64.26**	**73.33**	**61.93**	**86.70**	**1938.72**	**70.10**	
VisionZip+FT‡	97.38%	63.70	73.20	64.40	86.30	1829.00	66.60	
Reduction Rate ≈ 90%								
FaseterVLM	92.79%	59.68	71.24	60.14	80.39	1470.98	67.61	42.90
AdaV (Ours)	**95.40%**	**62.78**	**73.53**	**59.76**	**85.79**	**1603.05**	**69.67**	**39.70**
VisionZip	94.19%	60.70	70.30	60.90	82.00	1805.00	67.20	
AdaV (Ours)	**97.44%**	**62.78**	**73.53**	**59.76**	**85.79**	**1912.69**	**69.67**	
VisionZip+FT‡	96.90%	62.50	72.70	63.20	85.70	1861.00	66.90	
Reduction Rate ≈ 95%								
FaseterVLM	86.52%	56.14	70.40	58.43	73.81	1388.44	64.69	34.30
AdaV (Ours)	**93.14%**	**60.97**	**72.68**	**58.05**	**84.76**	**1557.43**	**68.56**	**37.90**
VisionZip	90.44%	57.80	69.30	58.40	76.60	1739.00	64.90	
AdaV (Ours)	**95.50%**	**60.97**	**72.68**	**58.05**	**84.76**	**1867.07**	**68.56**	
VisionZip+FT‡	93.89%	59.70	72.00	60.80	84.00	1766.00	65.30	

NEXT-34B, and LLaVA-v1.5-7B in Tab. A1, A2 and A3. The result demonstrates that the proposed AdaV achieves SOTA performance on various VLMs.

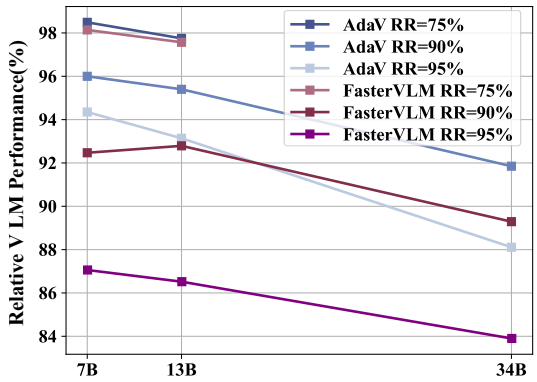

Figure A3: Influence of model scale for visual token pruning across token pruning methods.

Table A2: Comparison with SOTA approaches on LLaVA-NEXT-34B

Method	Average	GQA	SQA-IMG	TextVQA	POPE	MME	MMB	MM-Vet
LLaVA-NEXT-34B	100.00%	67.10	81.80	69.50	87.70	2028.00	79.30	57.40
Reduction Rate ≈ 90%								
FaseterVLM	89.29%	59.60	78.43	60.93	80.35	1869.73	75.85	42.00
AdaV (Ours)	**91.85%**	**62.71**	**79.08**	**57.92**	**86.67**	**1958.92**	**75.17**	**45.50**
Reduction Rate ≈ 95%								
FaseterVLM	83.90%	55.31	78.78	58.03	74.02	1745.38	71.64	36.90
AdaV (Ours)	**88.11%**	**60.12**	**78.43**	**55.05**	**86.44**	**1909.81**	**74.39**	**37.60**

Table A3: Comparison with SOTA approaches on LLaVA-v1.5-7B. † means that we report both the perception-only score and the summation of the perception score and the cognition score in parenthesis. ‡ with a gray background means the model is fine-tuned, which is expected to be stronger.

Method	Average	GQA	SQA-IMG	TextVQA	POPE	MME	MMB	MM-Vet
LLaVA-1.5-7B	100.00%	61.94	69.51	58.21	85.88	1506.47 (1862.00)	64.69	31.30
Reduction Rate 75%								
FastV	94.67%	56.58	69.11	57.38	73.74	1463.39	64.00	28.60
FitPrune	96.22%	59.38	69.01	56.49	80.75	1472.86	63.92	28.40
SparseVLM	93.22%	55.11	69.36	55.99	77.57	1351.65	59.54	29.90
FaseterVLM	98.32%	58.34	67.92	57.07	83.46	1433.76	62.54	34.20
AdaV (Ours)	**97.83%**	**58.38**	**69.31**	**56.66**	**84.72**	**1432.68**	**62.28**	**32.40**
VisionZip	96.12%	57.60	68.90	56.80	83.20	1761.70	62.00	30.00
AdaV (Ours)	**97.77%**	**58.38**	**69.31**	**56.66**	**84.72**	**1762.32**	**62.28**	**32.40**
VisionZip+FT‡	98.36%	58.90	68.30	57.00	83.70	1823.00	62.60	32.90
Reduction Rate 90%								
FastV	86.26%	51.20	69.81	54.75	57.30	1210.36	59.97	27.20
FitPrune	81.62%	49.96	68.22	56.49	53.81	1147.46	56.27	21.80
SparseVLM	78.87%	48.86	67.23	55.99	65.82	1030.61	49.05	18.60
FaseterVLM	92.91%	54.91	68.91	55.28	75.85	1348.63	60.57	30.10
AdaV (Ours)	**93.59%**	**55.30**	**68.82**	**54.53**	**82.33**	**1368.28**	**60.30**	**29.20**
VisionZip	94.02%	55.10	69.00	55.50	77.00	1690.00	60.10	31.70
AdaV (Ours)	**93.63%**	**55.30**	**68.82**	**54.53**	**82.33**	**1695.42**	**60.30**	**29.20**
VisionZip+FT‡	95.76%	58.90	68.80	56.00	80.90	1756.00	61.50	30.20
Reduction Rate 95%								
FastV	72.48%	46.03	70.00	51.56	35.47	971.56	50.17	18.90
FitPrune	65.85%	43.60	68.32	46.75	31.17	855.21	39.69	18.00
FaseterVLM	87.76%	51.51	69.56	53.09	67.24	1254.80	58.51	27.50
AdaV (Ours)	**88.32%**	**52.96**	**68.42**	**51.89**	**78.04**	**1313.36**	**58.51**	**24.00**