Learning to Predict Middle-Layer Attention in MLLMs for Visual Token Pruning

Yuyao Sun[1*], **Tao Deng**[1*], **Shuang Li**[1†], **Deqing Wang**[1†], **Hao Geng**[1], **Minjun Yu**[2]

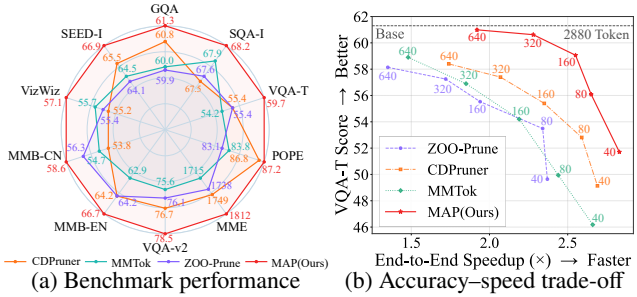
[1]Beihang University
[2]Shanghai EABOT Technology Co., Ltd.

Abstract

Multimodal large language models (MLLMs) achieve strong performance across diverse vision-language tasks, but their efficiency is limited by the cost of processing numerous visual tokens. Visual token pruning can reduce this cost, but requires accurate token importance estimates. Recent studies have demonstrated that text-to-vision attention from middle language model layers can effectively guide visual token pruning, typically using attention from a predefined middle layer to select the visual tokens to retain. Two problems therefore remain. First, our analysis shows that the layer whose attention is most responsive to the question varies substantially across samples, making a fixed layer suboptimal. Second, obtaining attention from the appropriate middle layer requires processing numerous visual tokens through several language model layers, by which point considerable computation has already been spent. To address both problems, we propose Middle-layer Attention Prediction (MAP), which uses Question Contrastive Teacher Selection to identify a sample-specific teacher layer by contrasting attention under the original and reference questions, and distills attention from the selected layer into a lightweight predictor that estimates visual token importance from multimodal input features. During inference, MAP combines the predicted importance scores with a diversity criterion to prune visual tokens before the first language model layer. Thus, MAP requires no attention maps for pruning and remains compatible with existing inference acceleration techniques. Across ten benchmarks on LLaVA-NeXT-7B, MAP retains 97.5% of the unpruned model's performance with only 5.56% of the visual tokens, yielding a 3.09× end-to-end speedup.

Introduction

Multimodal large language models (MLLMs) have demonstrated strong capabilities across diverse vision-language tasks (Alayrac et al. 2022; Li et al. 2023b; Dai et al. 2023; Liu et al. 2023, 2024a). However, processing visual inputs remains computationally expensive, as each image is typically represented by hundreds or even thousands of visual tokens. To preserve fine-grained visual information, recent MLLMs often represent visual inputs with substantially more visual tokens (Liu et al. 2024b; Li et al. 2024; Chen et al. 2024b; Wu et al. 2024; Zhu et al. 2025; Bai et al. 2025a), further increasing computational and memory costs. Consequently, visual token pruning has emerged as a promising approach to improving MLLM inference efficiency.



(a) Benchmark performance (b) Accuracy–speed trade-off

Figure 1: Performance and inference efficiency comparison on LLaVA-NeXT-7B. (a) Performance across ten vision-language benchmarks with 160 retained visual tokens. (b) TextVQA accuracy versus end-to-end speedup at different retention levels; numbers denote retained visual tokens.

A prominent line of work uses text-to-vision attention within the language model to guide visual token pruning (Chen et al. 2024a; Zhang et al. 2024; Liu et al. 2024c). However, recent studies have questioned whether attention scores reliably reflect token importance, showing that attention-guided pruning can even underperform random pruning at high compression ratios (Zhang et al. 2025a; Wen et al. 2025; Yang et al. 2025). LearnPruner (Takezoe et al. 2026) revisits this issue and shows that text-to-vision attention from middle language model layers can effectively guide visual token pruning. Specifically, it performs question-aware pruning in its second stage using attention extracted from a predefined middle layer. However, whether a fixed middle layer can provide effective pruning guidance across different inputs remains unclear.

To investigate this issue, we first analyze how text-to-vision attention localizes question-relevant visual evidence across language model layers. Using paired questions that refer to different objects in the same image, we find that attention from middle layers most clearly distinguishes regions relevant to the current question. More importantly, the layer providing the strongest localization varies substantially across samples. To assess whether these differences in localization are also reflected in pruning performance, we further evaluate layer-wise attention as a guidance signal for visual token pruning before LLM processing. We find that attention from middle layers

*These authors contributed equally.
†Corresponding author.

generally provides the strongest pruning guidance, while the best layer differs across benchmarks. These findings suggest that, although middle-layer attention is broadly effective for pruning, no single layer is universally optimal.

However, exploiting middle-layer attention poses a fundamental efficiency dilemma. Although it provides effective pruning guidance, obtaining the signal requires numerous visual tokens to pass through several language model layers, by which point a substantial portion of the computation has already been incurred. Moreover, explicitly materializing attention maps during inference can limit compatibility with optimized attention kernels, further reducing efficiency.

To address the variation in the optimal layer across samples and the cost of obtaining middle-layer attention, we propose **Middle-layer Attention Prediction (MAP)**. During offline training, **Question Contrastive Teacher Selection (QCTS)** compares text-to-vision attention distributions for the input question and a generic reference question on the same image. Since layers insensitive to the question tend to produce similar attention for both questions, QCTS selects the layer with the largest difference as the teacher for each sample. A lightweight predictor is then trained to match the selected attention distribution through distribution matching, while the original MLLM remains frozen throughout training.

During inference, the predictor estimates the sample-specific text-to-vision attention distribution from multimodal representations available before language model processing. MAP uses the predicted attention scores to assess visual token importance, combines these scores with a diversity criterion, and prunes tokens before the first language model layer, allowing only the retained tokens to participate in subsequent computation. By avoiding reliance on LLM attention maps during inference, MAP remains compatible with optimized attention kernels such as FlashAttention (Dao et al. 2022).

Extensive experiments across multiple MLLM backbones, benchmarks, and visual token retention ratios validate the effectiveness of MAP. MAP consistently achieves a better accuracy–efficiency trade-off than existing visual token pruning methods. Across ten benchmarks on LLaVA-NeXT-7B, MAP retains 97.5% performance using only 5.56% of visual tokens. At this retention ratio, MAP achieves a $7.44\times$ speedup in the LLM prefill stage and a $3.09\times$ speedup in end-to-end inference. Our contributions are summarized as follows:

- We show that middle-layer text-to-vision attention effectively guides pruning, while the layer most responsive to the question varies across samples.
- To avoid the cost of obtaining this attention during inference, we propose **MAP**, which distills it into a lightweight predictor operating before language model processing.
- We introduce QCTS, which contrasts attention distributions induced by the input and reference questions and selects the most question-responsive layer as a sample-specific teacher to supervise the predictor.
- Extensive experiments across MLLMs and benchmarks show that MAP achieves a better accuracy–efficiency trade-off than existing visual token pruning methods.

Related Work

Multimodal Large Language Models

Multimodal large language models (MLLMs) extend large language models to process visual inputs by connecting pre-trained vision encoders with language models through cross-modal interfaces, including cross-attention layers (Alayrac et al. 2022), Q-Former (Li et al. 2023b; Dai et al. 2023), and lightweight projectors (Liu et al. 2023, 2024a). To enhance visual perception and understanding, recent MLLMs encode visual inputs at a finer granularity to preserve detailed spatial and semantic information, producing long input sequences (Liu et al. 2024b; Li et al. 2024; Zhu et al. 2025; Guo et al. 2025; Bai et al. 2025a). Processing a large number of visual tokens introduces substantial computational overhead, leading to the development of visual token pruning methods.

Visual Token Pruning in MLLMs

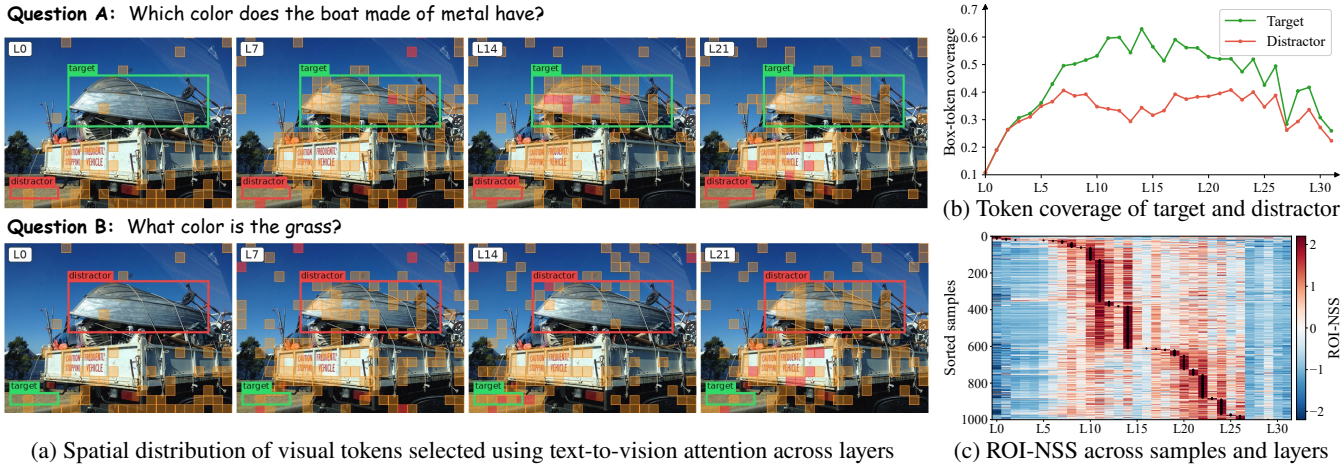
Visual token pruning improves MLLM efficiency by reducing the number of processed visual tokens while preserving performance. Many methods estimate token importance using attention scores (Chen et al. 2024a; Zhang et al. 2024; Ye et al. 2025), although recent studies question their reliability under aggressive compression (Zhang et al. 2025a; Wen et al. 2025). This has motivated alternative criteria such as instruction-conditioned diversity (Zhang et al. 2025b), multimodal coverage (Dong et al. 2025), and token sensitivity (Kim et al. 2026). LearnPruner (Takezoe et al. 2026) shows that text-to-vision attention is useful in middle layers but less effective in early and late layers, and uses a predefined middle layer for question-guided token selection. Extending this analysis, we identify two limitations: the most question-responsive layer varies across inputs, and middle-layer pruning still requires many visual tokens to traverse the preceding layers. MAP addresses both by using QCTS to select a sample-specific teacher layer through question contrast and distilling its attention into a predictor before the first LLM layer for efficient early pruning during inference.

Motivation

Question-Relevant Attention Across Layers

Effective pruning requires preserving visual evidence relevant to the question. We therefore examine whether middle-layer attention more clearly localizes question-relevant targets and whether the layer providing the strongest localization is consistent across inputs. To isolate question effects, we construct 500 grounded question pairs from GQA (Hudson and Manning 2019) (1,000 questions), each sharing an image but referring to different objects. For each question, the referred object is the target and the other distractor. We then analyze the visual tokens top-ranked by attention across layers.

Fig. 2(a) illustrates how attention localization changes across language model layers. At layer 0, token selection exhibits a strong positional bias, favoring tokens near the end of the sequence. At layer 7, the selections for the two questions remain similar despite their different targets. At layer 14, the selections become question-specific, focusing on the boat for Question A and the grass for Question B. This distinction weakens again in later layers.



(a) Spatial distribution of visual tokens selected using text-to-vision attention across layers

(b) Token coverage of target and distractor

(c) ROI-NSS across samples and layers

Figure 2: Analysis of how text-to-vision attention changes across language model layers. (a) The 128 visual tokens with the highest attention scores are highlighted. Target and distractor denote regions relevant and irrelevant to the question, respectively. (b) Average token coverage within the target and distractor boxes over 1000 questions. (c) ROI-NSS across questions and layers. ROI-NSS is the mean normalized attention score within the target box, with higher values indicating better localization.

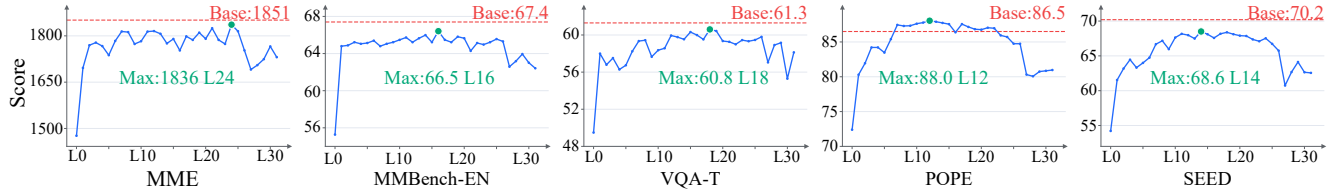


Figure 3: Layer-wise evaluation of text-to-vision attention for visual token pruning on LLaVA-NeXT-7B. We report performance after pre-LLM pruning, retaining the top 11.1% of visual tokens according to the attention scores from each layer.

To determine whether this pattern generalizes, Fig. 2(b) reports the average coverage of the target and distractor objects by the top 128 visual tokens at each layer. In early layers, their coverage is similar, indicating limited distinction between the two objects. Toward the middle layers, target coverage increases while distractor coverage decreases, showing clearer localization of evidence relevant to the current question. In later layers, coverage declines for both objects. Together, Fig. 2(a) and Fig. 2(b) show that middle-layer attention most clearly distinguishes the current target from the distractor, both qualitatively and across the evaluated questions.

Fig. 2(c) then examines whether the strongest localization occurs at the same layer across samples. ROI-NSS is computed by normalizing the attention scores across all visual tokens and averaging them within the target bounding box. For each question, the layer with the highest ROI-NSS is regarded as providing the strongest localization. Although higher ROI-NSS values are generally concentrated in the middle layers, the peak varies substantially across questions. Thus, no single predefined middle layer consistently provides the strongest question-relevant localization across inputs.

Pruning Utility of Text-to-Vision Attention

Having observed that the strongest question-relevant localization occurs at different layers across samples, we next show that the layer providing the best pruning guidance also differs across benchmarks. We evaluate this on five benchmarks with LLaVA-NeXT-7B by using attention from each language model layer to guide visual token pruning at layer 0. The unpruned model provides the baseline and layer-wise attention scores. Each layer's scores are then used in a separate forward pass that retains only the top 11.1% of visual tokens. Fig. 3 presents the results. Performance generally improves when attention is taken from early to middle layers and declines in later layers, while the best layer differs across benchmarks. With the best layer selected for each benchmark, the pruned model remains close to the unpruned baseline. These results show that middle-layer attention provides effective pruning guidance, but no single layer consistently performs best.

Method

Our analysis shows that middle-layer text-to-vision attention provides effective guidance for visual token pruning. However, the layer most responsive to the question varies across samples, making a fixed-layer strategy suboptimal. Moreover, extracting this signal requires numerous visual tokens to pass through several language model layers, limiting the achievable efficiency gains. To overcome both limitations, we propose MAP, which distills text-to-vision attention from a sample-specific teacher layer into a lightweight predictor operating before language model processing.

MAP consists of offline training and inference phases. During offline training, MAP selects the middle layer most responsive to the input question as the teacher for each sample and trains the predictor to reproduce its text-to-vision atten-

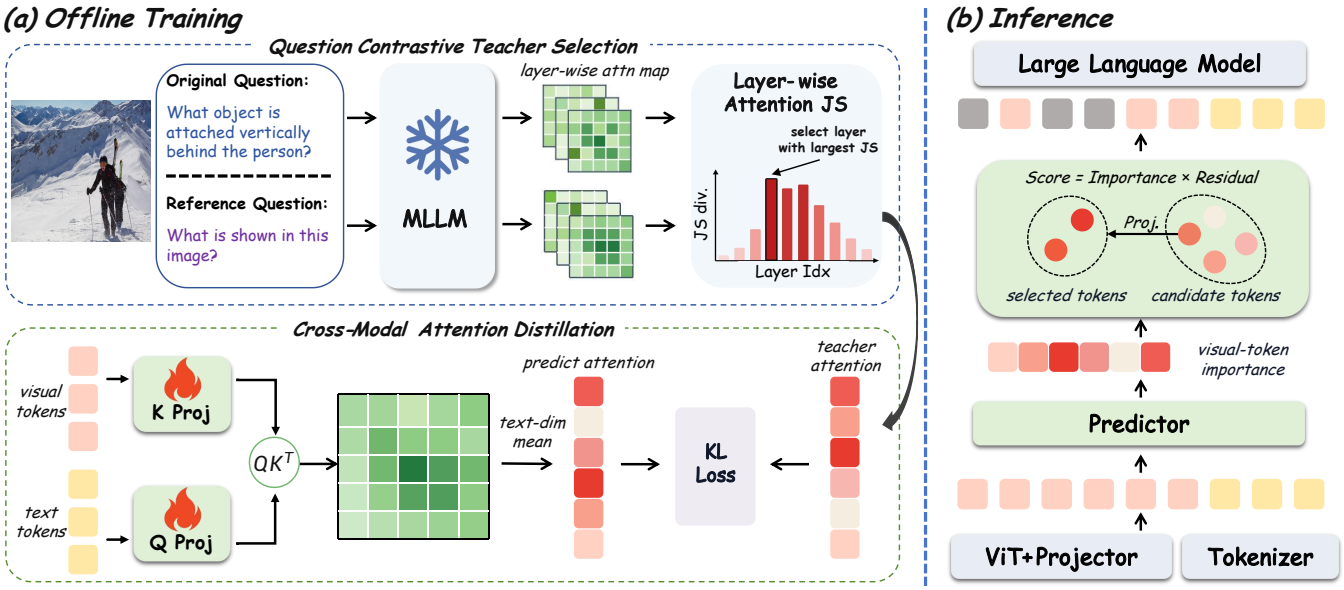


Figure 4: Overview of MAP.

tion distribution from pre-LLM multimodal representations. During inference, the predictor directly estimates text-to-vision attention. The predicted attention scores serve as visual token importance, which MAP combines with visual feature diversity to retain a compact token set balancing question relevance and complementary information.

Question Contrastive Teacher Selection

The utility of attention from different middle layers for visual token pruning varies across samples, making a fixed layer suboptimal as a universal teacher. To construct teacher supervision tailored to each sample, we propose Question Contrastive Teacher Selection (QCTS), which compares the text-to-vision attention produced for the input question with that produced for a generic reference question on the same image and then selects the layer with the largest discrepancy as the teacher. The selected teacher attention is expected to best capture visual regions relevant to the input question.

Given an image and question pair, the image is processed by the vision encoder and multimodal projector to produce N visual token embeddings $\mathbf{V} = [\mathbf{v}_1, \ldots, \mathbf{v}_N]$, while the M question tokens are mapped by the LLM embedding layer to text token embeddings $\mathbf{T} = [\mathbf{t}_1, \ldots, \mathbf{t}_M]$. Together, $\mathbf{V}$ and $\mathbf{T}$ constitute the multimodal input representations provided to the language model. Let $\mathcal{L} = \{0, \ldots, L-1\}$ denote the set of all L language model layers. At each layer $l \in \mathcal{L}$, we extract text-to-vision attention and average it over all H attention heads and M question tokens for each visual token:

$$a_n^l = \frac{1}{HM} \sum_{h=1}^{H} \sum_{m=1}^{M} A_{h,m,n}^l. \tag{1}$$

Here, $A_{h,m,n}^l$ denotes the attention from question token $\mathbf{t}_m$ to visual token $\mathbf{v}_n$ at head h of layer l. At each layer, we apply ℓ_1 normalization to the averaged attention scores over visual tokens, yielding the target attention distributions $\{\mathbf{p}^l\}_{l \in \mathcal{L}}$.

We obtain the reference attention distributions by pairing the same image with the fixed reference question, "What is shown in this image?", while keeping the remaining input unchanged. Applying the attention aggregation and normalization procedure produces $\mathbf{p}_{\text{ref}}^l$ at each layer, with attention averaged over the M_{ref} reference question tokens. As the reference question does not specify a particular object, attribute, relation, or region, it provides a generic visual attention baseline. The discrepancy between $\mathbf{p}^l$ and $\mathbf{p}_{\text{ref}}^l$ therefore quantifies how the input question redirects visual attention relative to this baseline at layer l. Specifically, we quantify the discrepancy at each layer using the Jensen–Shannon divergence and select the layer with the largest value:

$$l^\star = \arg\max_{l \in \mathcal{L}} D_{\text{JS}} \left(\mathbf{p}^l, \mathbf{p}_{\text{ref}}^l \right). \tag{2}$$

The attention distribution $\mathbf{p}^{l^\star}$ from the selected layer is then used as supervision to train the predictor.

Cross-Modal Attention Distillation

After QCTS selects the teacher distribution $\mathbf{p}^{l^\star}$ for each training sample, we train a lightweight predictor to reproduce it directly from the multimodal input representations of the LLM. The predictor first transforms the text and visual representations using two modality-specific MLPs, f_T and f_V, and then projects them into queries and keys using $\mathbf{W}_Q$ and $\mathbf{W}_K$, respectively. Before computing the attention scores, we apply RoPE to the projected queries and keys to preserve positional information. The predicted attention distribution over visual tokens is computed as

$$\hat{p}_n = \frac{1}{M} \sum_{m=1}^{M} \text{softmax}_n \left(\frac{\langle \mathbf{W}_Q f_T(\mathbf{t}_m), \mathbf{W}_K f_V(\mathbf{v}_n) \rangle}{\sqrt{d}} \right), \tag{3}$$

where d is the projection dimension. Since only attention weights are required, value and output projections are omitted.

Method	GQA	SQA^{I}	VQA^{T}	POPE	MME	VQA^{v2}	MMB^{EN}	MMB^{CN}	VizWiz	SEED-I	Rel.↑
All 576 Tokens (100.0%)											
Vanilla	61.9	69.5	58.2	85.9	1862	78.5	64.7	58.3	50.0	66.1	100.0%
Retain 128 Tokens (↓ 77.8%)											
VisionZip (CVPR2025)	57.6	68.7	56.9	83.3	1761	75.6	62.1	57.0	51.6	61.2	96.7%
DART (EMNLP2025)	57.9	69.1	56.3	80.4	1721	74.7	60.7	57.3	52.8	62.2	96.3%
CDPruner (NeurIPS2025)	59.9	68.6	56.2	87.4	1745	76.6	63.1	55.0	52.8	63.2	97.8%
MMTok (ICLR2026)	59.2	68.8	57.0	86.3	1779	76.3	62.3	55.1	52.9	63.1	97.8%
ZOO-Prune (CVPR2026)	59.4	68.9	57.8	87.1	1751	76.5	61.8	55.6	52.1	62.8	97.7%
MAP (Ours)	60.1	68.7	57.7	87.3	1800	76.9	63.1	57.0	52.5	64.4	**98.9%**
Retain 64 Tokens (↓ 88.9%)											
VisionZip (CVPR2025)	55.1	69.0	55.5	77.0	1690	72.4	60.1	55.4	52.9	57.3	93.7%
DART (EMNLP2025)	54.7	69.3	54.7	73.8	1705	71.3	48.0	53.5	53.5	59.6	91.3%
CDPruner (NeurIPS2025)	58.6	68.1	55.3	87.1	1717	75.4	61.1	53.2	53.4	62.1	96.4%
MMTok (ICLR2026)	58.2	69.1	56.0	85.7	1715	75.2	61.2	54.5	53.9	61.4	96.6%
LearnPruner (ICLR2026)	58.9	68.3	56.6	86.8	1750	76.0	62.6	55.7	–	–	96.9%
ZOO-Prune (CVPR2026)	58.5	68.2	55.3	85.8	1657	75.0	60.2	54.7	54.0	60.9	95.9%
MAP (Ours)	59.4	68.9	57.0	87.1	1744	76.1	63.2	57.0	52.8	62.9	**98.1%**
Retain 32 Tokens (↓ 94.4%)											
VisionZip (CVPR2025)	51.8	69.1	53.1	69.4	1536	67.1	57.0	50.3	52.4	53.1	88.3%
DART (EMNLP2025)	52.9	69.3	52.2	69.1	1615	67.1	58.5	50.0	52.5	58.8	89.8%
CDPruner (NeurIPS2025)	57.0	69.0	53.2	87.4	1657	73.6	59.6	49.6	53.1	60.8	94.3%
MMTok (ICLR2026)	56.6	69.0	53.4	85.0	1635	73.4	59.4	49.8	54.6	59.7	93.9%
LearnPruner (ICLR2026)	57.2	68.2	56.1	84.5	1672	74.0	60.8	55.5	–	–	94.8%
ZOO-Prune (CVPR2026)	55.8	68.7	54.1	84.7	1643	72.2	58.9	50.7	53.7	58.2	93.4%
MAP (Ours)	58.1	69.1	55.3	87.4	1726	74.6	62.2	54.6	52.9	61.5	**96.6%**

Table 1: Performance comparison on LLaVA-1.5-7B with different numbers of retained visual tokens. Rel. denotes the average performance across benchmarks after normalizing each score by the corresponding vanilla score.

During offline training, the predictor takes the multimodal input embeddings as input and is optimized by minimizing $D_{\mathrm{KL}}(\mathbf{p}^{l^\star} \| \widehat{\mathbf{p}})$. This objective distills the visual token importance from the selected teacher distribution into the predictor.

Inference-Time Token Pruning

During inference, MAP treats the predicted attention $\widehat{\mathbf{p}}$ as visual token importance. Given a target number K of visual tokens to retain, MAP first keeps the top K_0 tokens with the highest predicted attention to form the initial retained set $\mathcal{R}_0$, while the remaining tokens form the candidate set $\mathcal{C}_0$. Starting from this initialization, MAP greedily adds one candidate token at each iteration until K tokens are retained, jointly considering predicted importance and feature diversity.

To promote diversity, MAP maintains an orthonormal basis constructed from the selected visual embeddings. The initial basis $\mathbf{B}_0$ is obtained by orthonormalizing the embeddings in $\mathcal{R}_0$ and is updated whenever a new token is retained. At iteration t, the next token is selected by

$$i_t = \underset{i \in \mathcal{C}_{t-1}}{\arg\max}\, \widehat{p}_i \left\| \mathbf{v}_i - \mathbf{B}_{t-1}\mathbf{B}_{t-1}^{\top}\mathbf{v}_i \right\|_2 . \tag{4}$$

The importance term favors tokens related to the input question, while the orthogonal residual rewards features complementary to the current subspace. After i_t is selected, it is moved from the candidate set to the retained set. Its embedding $\mathbf{v}_{i_t}$ is then orthogonalized against $\mathbf{B}_{t-1}$, and the normalized residual is appended to form $\mathbf{B}_t$. This procedure continues until the retained set contains K tokens. Before language model processing, the selected tokens are restored to their original order while preserving their positional indices and are then combined with the question tokens.

Experiments

Datasets and Models. To empirically validate the effectiveness and generalizability of MAP, we conduct comprehensive experiments on three widely used MLLM backbones: LLaVA-1.5-7B (Liu et al. 2024a), LLaVA-NeXT-7B (Liu et al. 2024b), and Qwen2.5-VL-7B-Instruct (Bai et al. 2025b). For LLaVA-1.5-7B and LLaVA-NeXT-7B, we evaluate MAP on GQA (Hudson and Manning 2019), SQA (Lu et al. 2022), TextVQA (Singh et al. 2019), POPE (Li et al. 2023c), MME (Fu et al. 2023), VQAv2 (Goyal et al. 2017), MMBench (Liu et al. 2024d), VizWiz (Gurari et al. 2018), and SEED-Bench (Li et al. 2023a). For Qwen2.5-VL-7B-Instruct, we report results on AI2D (Kembhavi et al. 2016), MME, MMBench, TextVQA, and GQA.

Compared Methods. We compare MAP with representative token pruning methods. FastV (Chen et al. 2024a) prunes visual tokens using early layer attention. VisionZip (Yang et al. 2025) selects tokens using ViT attention. DART (Wen et al. 2025) removes redundant tokens by similarity to pivot tokens. CDPruner (Zhang et al. 2025b) balances relevance and diversity with an instruction-conditioned DPP. MMTok (Dong et al. 2025) formulates token selection as a multimodal maximum coverage problem. ZOO-Prune (Kim et al. 2026) estimates token importance through sensitivity analysis. LearnPruner (Takezoe et al. 2026) removes visual redundancy using learned importance and diversity, then prunes query-irrelevant tokens based on middle-layer text-to-vision attention.

Main Results.

Tables 1 and 2 report comparisons on LLaVA-1.5-7B and LLaVA-NeXT-7B, respectively. MAP achieves the best av-

Method	GQA	SQA^{I}	VQA^{T}	POPE	MME	VQA^{v2}	MMB^{EN}	MMB^{CN}	VizWiz	SEED-I	Rel.↑
Upper Bound, All 2880 Tokens (100.0%)											
Vanilla	64.2	70.1	61.3	86.5	1851	81.8	67.4	60.6	57.6	70.2	100.0%
Retain 640 Tokens (↓ 77.8%)											
VisionZip (CVPR2025)	61.2	68.1	59.9	86.0	1787	79.1	65.8	58.1	57.1	66.7	97.1%
DART (EMNLP2025)	61.3	68.2	59.5	85.0	1721	78.3	64.9	57.1	57.0	67.9	96.3%
CDPruner (NeurIPS2025)	62.6	67.9	58.4	87.3	1800	79.9	66.3	57.5	55.6	68.6	97.3%
MMTok (ICLR2026)	62.2	68.4	58.9	86.7	1829	79.3	65.2	56.5	55.8	67.7	97.0%
ZOO-Prune (CVPR2026)	62.2	67.7	58.0	86.7	1783	79.6	65.2	57.2	55.2	67.9	96.6%
MAP (Ours)	62.8	68.9	61.0	88.2	1838	80.4	68.4	59.6	57.6	69.5	**99.4%**
Retain 320 Tokens (↓ 88.9%)											
VisionZip (CVPR2025)	58.9	67.5	58.8	82.5	1702	76.2	63.3	55.6	56.2	63.4	93.8%
DART (EMNLP2025)	59.5	67.5	57.6	81.0	1705	75.7	64.2	55.7	56.8	64.8	93.9%
CDPruner (NeurIPS2025)	61.6	67.8	57.4	87.2	1807	78.4	65.5	55.7	55.8	67.1	96.2%
MMTok (ICLR2026)	60.9	67.3	56.9	85.7	1799	77.6	64.3	55.8	55.4	66.2	95.3%
LearnPruner (ICLR2026)	62.2	68.6	58.4	–	1845	78.3	66.8	–	–	–	97.4%
ZOO-Prune (CVPR2026)	61.0	67.2	57.3	85.5	1787	78.1	64.9	56.4	55.1	66.5	95.5%
MAP (Ours)	61.8	68.6	60.6	87.8	1842	79.6	67.1	59.0	57.5	68.4	**98.5%**
Retain 160 Tokens (↓ 94.4%)											
VisionZip (CVPR2025)	55.2	67.9	55.0	75.8	1630	71.4	58.6	50.4	55.5	58.3	88.5%
DART (EMNLP2025)	56.8	67.8	54.9	75.3	1615	72.5	62.0	53.6	56.7	60.6	90.3%
CDPruner (NeurIPS2025)	60.8	67.5	55.4	86.8	1749	76.7	64.2	53.8	55.2	65.5	94.3%
MMTok (ICLR2026)	60.0	67.9	54.2	83.8	1715	75.6	62.9	54.7	55.7	64.5	93.3%
LearnPruner (ICLR2026)	58.7	67.6	55.0	–	1784	76.2	65.3	–	–	–	94.0%
ZOO-Prune (CVPR2026)	59.9	67.6	55.4	83.1	1738	76.1	64.2	56.3	55.4	64.1	93.9%
MAP (Ours)	61.3	68.2	59.7	87.2	1812	78.5	66.7	58.6	57.1	66.9	**97.5%**

Table 2: Performance comparison on LLaVA-NeXT-7B with different numbers of retained visual tokens.

Method	AI2D	MME	MMB^{EN}	MMB^{CN}	VQA^{T}	GQA	Rel.↑
All 1280 Tokens (100.0%)							
Vanilla	83.2	2325	83.7	80.5	83.0	61.1	100.0%
Retain 512 Tokens (↓ 60%)							
FastV	78.8	2317	81.5	79.1	81.4	58.4	97.3%
VisionZip	82.7	2337	81.8	79.7	80.8	60.1	98.7%
MAP	82.8	2350	82.9	79.8	82.2	60.8	**99.6%**
Retain 256 Tokens (↓ 80%)							
FastV	76.2	2238	78.8	74.4	76.2	53.6	92.3%
VisionZip	79.3	2209	79.7	76.6	74.2	57.3	94.0%
MAP	81.0	2348	82.2	78.0	79.7	59.9	**97.9%**
Retain 128 Tokens (↓ 90%)							
FastV	68.2	1745	69.5	65.3	58.3	52.1	79.4%
VisionZip	72.5	2168	76.0	74.5	60.4	54.0	87.5%
MAP	79.4	2298	80.3	76.3	75.1	58.6	**95.2%**

Table 3: Performance comparison on Qwen2.5-VL-7B.

erage performance across all settings on both backbones. When retaining only **22.2%**, **11.1%**, and **5.6%** of the original visual tokens, MAP outperforms the strongest baselines by **1.1**, **1.2**, and **1.8** points on LLaVA-1.5-7B, and by **2.1**, **1.1**, and **3.2** points on LLaVA-NeXT-7B. The advantage becomes pronounced as the retention ratio decreases, indicating that MAP preserves question-relevant visual evidence more effectively under aggressive pruning. Its consistent gains on LLaVA-NeXT-7B demonstrate its effectiveness on longer visual sequences, where token selection is more challenging.

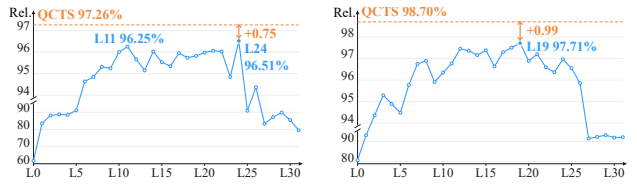
We further evaluate MAP on Qwen2.5-VL-7B-Instruct in Table 3. MAP consistently outperforms existing methods across all token retention ratios, improving the average relative score over the strongest baseline by **0.9** points with 512 tokens, **3.9** points with 256 tokens, and a substantially larger **7.7** points with 128 tokens. Notably, when restricted to only 128 tokens, existing methods degrade sharply, whereas MAP maintains an average relative score of **95.2%**. These results demonstrate that MAP better preserves question-relevant visual evidence at high compression ratios.

Method	Cov.↑	ROI-NSS↑	Layer Dist.↓	Hit@1↑	Hit@3↑	Hit@5↑
ROI-Max	66.2%	2.728	0.0	100.0%	100.0%	100.0%
Random	56.5%	1.210	6.5	5.8%	16.0%	26.3%
Layer 14	61.3%	2.056	4.6	22.8%	42.5%	52.7%
QCTS	**64.1%**	**2.421**	**3.3**	**47.8%**	**69.8%**	**79.8%**

Table 4: Comparison of layer-selection strategies.

Ablation of Teacher Selection for Direct Pruning

QCTS is designed to identify a teacher layer whose attention best captures the visual regions relevant to each question. To assess this ability, we use the same 1,000 image–question samples from the preceding analysis. ROI-Max selects the layer with the highest ROI-NSS for each question. Random uniformly selects one layer from Layers 8–24 for each question, while Layer 14 is the best-performing fixed baseline. For each method, we retain the top 128 visual tokens according to the attention scores from its identified layer. Macro Coverage measures the average target-region token coverage, ROI-NSS measures the concentration of attention within the target region, Layer Dist. and Hit@K measure how closely the identified layer matches ROI-Max. As shown in Table 4, QCTS consistently outperforms both baselines, achieving 64.1% Macro Coverage, an ROI-NSS of 2.421, and a Layer Dist. of 3.3. Its Hit@1/3/5 rates reach 47.8%, 69.8%, and 79.8%, respectively. The coverage and ROI-NSS demonstrate



(a) LLaVA-1.5-7B (b) LLaVA-NeXT-7B

Figure 5: Pruning performance of different layers and QCTS.

Teacher Attn	GQA	TextVQA	POPE	MME	MMB$^{EN/CN}$	Rel.
Vanilla	61.9	58.2	85.9	1862	64.7/58.3	100.0%
Single-layer teacher selection						
Fixed Layer 11	58.7	55.9	87.1	1757	62.5/56.3	96.6%
Fixed Layer 14	59.0	56.9	86.8	1698	62.7/56.2	96.4%
Fixed Layer 24	58.7	56.6	86.3	1745	63.1/55.4	96.5%
QCTS	59.4	57.0	87.1	1744	63.2/57.0	97.4%
Multi-layer teacher fusion						
QCTS-2	58.8	56.8	87.1	1735	62.6/56.3	96.7%
QCTS-4	58.9	56.7	86.9	1737	63.0/56.4	96.9%
QCTS-8	59.1	56.5	87.3	1739	63.2/56.0	96.9%
QCTS-16	58.5	56.3	87.2	1726	62.5/55.9	96.3%

Table 5: Performance of predictors trained with different teacher-selection strategies on LLaVA-1.5-7B at 64 tokens.

that QCTS preserves question-relevant visual information, while the low Layer Dist. and high Hit@K rates show that it reliably identifies layers close to ROI-Max.

We further evaluate whether the layers selected by QCTS provide effective signal for visual token pruning. Figure 5 compares it with attention from each fixed layer when only the top 11.1% of visual tokens ranked by text-to-vision attention are retained. The y-axis represents the average relative performance across the six benchmarks in Table 5. QCTS achieves average relative performance of 97.26% and 98.70% on LLaVA-1.5-7B and LLaVA-NeXT-7B, exceeding the strongest fixed-layer results by 0.75 and 0.99 percentage points, respectively. These results show that attention from the layers identified by QCTS is more effective for visual token pruning than attention from any fixed layer.

Ablation of Teacher Selection for Predictor Training

We investigate the effect of teacher layer selection by training predictors with different strategies and comparing their pruning performance. A fixed-layer teacher uses attention from the same predefined layer as supervision for all training samples. We report Layers 11, 14, and 24, as they are the three best performing fixed layer teachers. As shown in Table 5, QCTS achieves the best overall pruning performance, with a relative score of 97.4%, surpassing the strongest fixed-layer teacher by 0.8 points. This improvement suggests that QCTS can adaptively select a more suitable teacher layer for each sample, thereby providing more effective supervision for predictor training. We further evaluate QCTS-n, which averages the attention from the top n layers selected by QCTS for each sample as teacher supervision. However, multi-layer fusion provides no additional gains and instead leads to a slight decrease in pruning performance. These results show that QCTS outperforms both fixed-layer selection and multi-layer teacher fusion. We therefore use QCTS to select the teacher layer for training the predictor in all experiments.

Selection	Tokens	GQA	TextVQA	POPE	MME	MMB$^{EN/CN}$	Rel.↑
Top-K	128	60.0	57.3	87.1	1789	63.3/56.8	98.0%
+Diversity	128	60.1	57.7	87.3	1800	63.1/57.0	98.3%
Top-K	64	59.0	56.4	87.1	1721	63.7/56.4	96.9%
+Diversity	64	59.4	57.0	87.1	1744	63.2/57.0	97.4%
Top-K	32	57.1	54.3	87.4	1642	62.8/53.9	94.2%
+Diversity	32	58.1	55.3	87.4	1726	62.2/54.6	95.5%

Table 6: Effect of diversity-aware selection with different numbers of retained visual tokens on LLaVA-1.5-7B.

Method	Tokens	Prefill (ms)	End-to-End (ms)	KV Cache (MB)
LLaVA-1.5-7B				
Vanilla	576	58.9	101.5	336
MAP	32	21.5 (2.73×)	67.2 (1.51×)	64 (-80.95%)
LLaVA-NeXT-7B				
Vanilla	2880	224.5	277.9	1488
MAP	160	30.2 (7.44×)	89.8 (3.09×)	128 (-91.39%)

Table 7: Average inference efficiency across ten benchmarks.

Effect of Diversity-Aware Token Selection

Table 6 studies the contribution of diversity-aware selection under different token budgets. The Top-K baseline directly retains the K tokens with the highest predicted attention scores, whereas the diversity-aware variant additionally considers feature complementarity. At 128 tokens, diversity-aware selection provides only a marginal improvement, from 98.0% to 98.3%. This result indicates that the predicted importance scores accurately identify the question-relevant tokens and already cover the key visual evidence under a sufficient token budget, leaving limited room for diversity to help. As the budget decreases, however, the improvement grows from 96.9% to 97.4% at 64 tokens and from 94.2% to 95.5% at 32 tokens. Under aggressive pruning, redundancy among highly ranked tokens becomes more costly. Diversity-aware selection mitigates this issue by preserving complementary visual features, leading to larger gains at smaller token budgets.

Efficiency Analysis

Table 7 reports average efficiency across ten benchmarks. MAP achieves up to 2.73× prefill speedup and 80.95% KV cache reduction on LLaVA-1.5-7B, and 7.44× prefill and 3.09× end-to-end speedups with 91.39% KV cache reduction on LLaVA-NeXT-7B. Detailed results for all ten benchmarks and a runtime breakdown are provided in the appendix.

Conclusion

In this work, we show that the middle layer most responsive to the question varies across samples and its attention is costly to obtain. MAP addresses both issues by using QCTS to select a sample-specific teacher and distilling its attention into a lightweight predictor. At inference, MAP combines predicted importance with feature diversity to prune visual tokens before the first LLM layer, without requiring attention maps. Experiments on three MLLM backbones demonstrate the favorable accuracy–efficiency trade-off of MAP.

References

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Bai, S.; Cai, Y.; Chen, R.; Chen, K.; Chen, X.; Cheng, Z.; Deng, L.; Ding, W.; Gao, C.; Ge, C.; et al. 2025a. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025b. Qwen2.5-VL Technical Report. *arXiv preprint arXiv:2502.13923*.

Chen, L.; Zhao, H.; Liu, T.; Bai, S.; Lin, J.; Zhou, C.; and Chang, B. 2024a. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *ECCV*.

Chen, Z.; Wang, W.; Tian, H.; Ye, S.; Gao, Z.; Cui, E.; Tong, W.; Hu, K.; Luo, J.; Ma, Z.; et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*.

Dai, W.; Li, J.; Li, D.; Tiong, A.; Zhao, J.; Wang, W.; Li, B.; Fung, P. N.; and Hoi, S. 2023. Instructblip: Towards general-purpose vision-language models with instruction tuning. In *NeurIPS*.

Dao, T.; Fu, D.; Ermon, S.; Rudra, A.; and Ré, C. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *NeurIPS*.

Dong, S.; Hu, J.; Zhang, M.; Yin, M.; Fu, Y.; and Qian, Q. 2025. Mmtok: Multimodal coverage maximization for efficient inference of vlms. *arXiv preprint arXiv:2508.18264*.

Fu, C.; Chen, P.; Shen, Y.; Qin, Y.; Zhang, M.; Lin, X.; Yang, J.; Zheng, X.; Li, K.; Sun, X.; et al. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*.

Goyal, Y.; Khot, T.; Summers-Stay, D.; Batra, D.; and Parikh, D. 2017. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *CVPR*.

Guo, D.; Wu, F.; Zhu, F.; Leng, F.; Shi, G.; Chen, H.; Fan, H.; Wang, J.; Jiang, J.; Wang, J.; et al. 2025. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Gurari, D.; Li, Q.; Stangl, A. J.; Guo, A.; Lin, C.; Grauman, K.; Luo, J.; and Bigham, J. P. 2018. Vizwiz grand challenge: Answering visual questions from blind people. In *CVPR*.

Hudson, D. A.; and Manning, C. D. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *CVPR*.

Kembhavi, A.; Salvato, M.; Kolve, E.; Seo, M.; Hajishirzi, H.; and Farhadi, A. 2016. A diagram is worth a dozen images. In *ECCV*.

Kim, Y.; Zhang, Y.; Liu, H.; Jung, A.; Lee, S.; and Hong, S. 2026. ZOO-Prune: Training-Free Token Pruning via Zeroth-Order Gradient Estimation in Vision-Language Models. In *CVPR*.

Li, B.; Wang, R.; Wang, G.; Ge, Y.; Ge, Y.; and Shan, Y. 2023a. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023b. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *ICML*.

Li, Y.; Du, Y.; Zhou, K.; Wang, J.; Zhao, X.; and Wen, J.-R. 2023c. Evaluating object hallucination in large vision-language models. In *EMNLP*.

Liu, H.; Li, C.; Li, Y.; and Lee, Y. J. 2024a. Improved baselines with visual instruction tuning. In *CVPR*.

Liu, H.; Li, C.; Li, Y.; Li, B.; Zhang, Y.; Shen, S.; and Lee, Y. J. 2024b. Llava-next: Improved reasoning, ocr, and world knowledge.

Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual instruction tuning. In *NeurIPS*.

Liu, T.; Shi, L.; Hong, R.; Hu, Y.; Yin, Q.; and Zhang, L. 2024c. Multi-Stage Vision Token Dropping: Towards Efficient Multimodal Large Language Model. *arXiv preprint arXiv:2411.10803*.

Liu, Y.; Duan, H.; Zhang, Y.; Li, B.; Zhang, S.; Zhao, W.; Yuan, Y.; Wang, J.; He, C.; Liu, Z.; et al. 2024d. Mmbench: Is your multi-modal model an all-around player? In *ECCV*.

Lu, P.; Mishra, S.; Xia, T.; Qiu, L.; Chang, K.-W.; Zhu, S.-C.; Tafjord, O.; Clark, P.; and Kalyan, A. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*.

Singh, A.; Natarajan, V.; Shah, M.; Jiang, Y.; Chen, X.; Batra, D.; Parikh, D.; and Rohrbach, M. 2019. Towards vqa models that can read. In *CVPR*.

Takezoe, R.; Li, Y.; Bo, Z.; Hou, A.; Guang, M.; and Long, K. 2026. LearnPruner: Rethinking Attention-based Token Pruning in Vision Language Models. *arXiv preprint arXiv:2604.23950*.

Wen, Z.; Gao, Y.; Wang, S.; Zhang, J.; Zhang, Q.; Li, W.; He, C.; and Zhang, L. 2025. Stop Looking for "Important Tokens" in Multimodal Language Models: Duplication Matters More. In *EMNLP*.

Wu, Z.; Chen, X.; Pan, Z.; Liu, X.; Liu, W.; Dai, D.; Gao, H.; Ma, Y.; Wu, C.; Wang, B.; et al. 2024. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*.

Yang, S.; Chen, Y.; Tian, Z.; Wang, C.; Li, J.; Yu, B.; and Jia, J. 2025. Visionzip: Longer is better but not necessary in vision language models. In *CVPR*.

Ye, W.; Wu, Q.; Lin, W.; and Zhou, Y. 2025. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *AAAI*.

Zhang, Q.; Cheng, A.; Lu, M.; Zhang, R.; Zhuo, Z.; Cao, J.; Guo, S.; She, Q.; and Zhang, S. 2025a. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *ICCV*.

Zhang, Q.; Liu, M.; Li, L.; Lu, M.; Zhang, Y.; Pan, J.; She, Q.; and Zhang, S. 2025b. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. In *NeurIPS*.

Zhang, Y.; Fan, C.-K.; Ma, J.; Zheng, W.; Huang, T.; Cheng, K.; Gudovskiy, D.; Okuno, T.; Nakata, Y.; Keutzer, K.; et al. 2024. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Tian, H.; Duan, Y.; Su, W.; Shao, J.; et al. 2025. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.