Spectral Evolution-Guided Token Pruning in Multimodal Large Language Models

Bin Chen[1,2], Yuxiang Cai*[1,2], Yadan Luo[3], Yi Zhang[4], Jianwei Yin[1,2], and Zhi Chen*[5]⋆

[1] School of Software Technology, Zhejiang University, Ningbo, China
[2] Zhejiang Key Laboratory of Digital-Intelligence Service Technology, China
[3] The University of Queensland, St Lucia, QLD, Australia
[4] Singapore Management University, Singapore
[5] The University of Southern Queensland, Toowoomba, QLD, Australia
caiyuxiang@zju.edu.cn, zhi.chen@unisq.edu.au
https://github.com/zjubinchen/CLSE

Abstract. Reducing visual token redundancy is critical for accelerating Multimodal Large Language Models (MLLMs) without degrading cross-modal reasoning performance. Existing token pruning methods typically rely on single-layer signals, such as attention scores or token similarities, which overlook the cross-layer transformation of visual representations and may exhibit positional bias in multimodal token sequences. To address this limitation, we propose a training-free token pruning framework based on Cross-Layer Spectral Evolution (CLSE). Instead of measuring token importance from single-layer feature magnitudes, CLSE quantifies how token representations evolve across Transformer layers in the frequency domain. This evolution reflects the transition from high-frequency structural details to low-frequency semantic abstractions. We observe that tokens with stronger spectral redistribution across layers are more likely to be semantically active and should therefore be preserved. By modeling cross-layer token dynamics, CLSE provides a stable importance criterion that mitigates positional bias. Extensive experiments on both image and video benchmarks demonstrate that CLSE achieves a superior trade-off between efficiency and accuracy under aggressive token reduction. Across multiple MLLMs, CLSE reduces FLOPs, KV cache memory, and latency while maintaining competitive or improved performance.

Keywords: MLLMs · Spectral Evolution · Token Pruning

1 Introduction

Multimodal Large Language Models (MLLM) [30,53] have achieved remarkable progress in visual question answering, grounded dialogue, and multimodal reasoning. By coupling a vision encoder (*e.g.,* ViT [8]) with a powerful LLM, these

⋆ Corresponding Authors: Yuxiang Cai, Zhi Chen

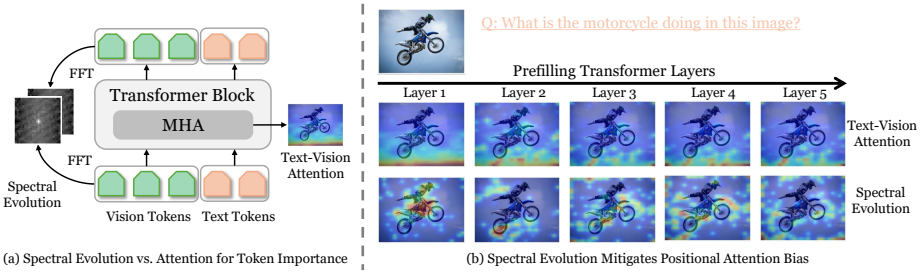


Fig. 1: Spectral evolution provides a more reliable token importance signal than attention. (a) Comparison between conventional text–vision attention and the proposed spectral evolution for estimating visual token importance across a Transformer block. (b) Layer-wise visualization shows that the attention is biased toward the visual tokens appearing later in the flattened sequence due to positional encoding effects, whereas spectral evolution consistently highlights semantically evolving regions across layers.

models enable strong cross-modal alignment and reasoning capabilities. However, this performance comes at a significant computational cost. Modern LLMs process hundreds to thousands of visual tokens per sample, leading to substantial inference latency, particularly in high-resolution and video scenarios.

A key source of inefficiency lies in visual token redundancy. Spatial patches in images and even temporal patches in videos exhibit strong local and cross-frame correlations. Yet standard Transformer architectures process all tokens uniformly through deep stacks of self-attention layers. To mitigate this inefficiency, recent works propose dynamic token reduction strategies. For example, DynamicViT [36], A-ViT [57], and ATS [10] introduce adaptive token pruning mechanisms in ViTs. In the multimodal domain, FastV [5] prunes visual tokens after early layers of MLLMs, while SparseVLM [64] adopts progressive sparsification with token recycling. Complementary approaches such as ToMe [3] and PruMerge [37] aggregate similar tokens instead of discarding them, preserving complementary information under aggressive compression.

Despite the effectiveness, most existing methods rely on *instantaneous* signals computed within a single layer, such as attention weights, feature magnitudes, or similarity scores. However, attention is not always a reliable indicator of semantic contribution. As illustrated in Fig. 1, attention scores can exhibit strong positional bias, *i.e.,* concentrating on visual tokens appearing later in the flattened sequence, even when those regions are not semantically critical [48]. Furthermore, static magnitude-based methods overlook a fundamental characteristic of LLM decoding behavior in an MLLM, where visual token representations undergo structured evolution across LLM layers, transitioning from low-level spatial details to high-level semantic abstractions while the text instruction is fused. This cross-layer transformation, rather than instantaneous attention intensity, more faithfully reflects a token's true contribution to multimodal reasoning.

Recent studies provide a complementary perspective by analyzing Transformers in the frequency domain. Works such as FNet [23] and Fourier-based vision architectures [35] demonstrate that spectral operations can effectively model global dependencies. More recently, frequency-aware compression strategies have been explored for multimodal models [43], highlighting the role of low-frequency components in semantic representation. Empirical observations suggest that early Transformer layers preserve high-frequency spatial information, while deeper layers progressively emphasize low-frequency semantic structures. This depth-wise spectral shift reflects the abstraction process underlying multimodal reasoning.

Motivated by these observations, we propose a cross-layer spectral evolution method for visual token reduction in MLLMs, as shown in Fig. 1 (a). Our key insight is that token importance should be measured by how its representation *evolves* across layers. The tokens that contribute meaningfully to text-vision reasoning show stronger spectral evolution during the detail-to-semantics transformation, whereas redundant background tokens remain relatively stationary in the frequency domain. Specifically, we (1) project visual tokens into the frequency domain and apply a Gaussian high-pass filter to isolate global information; (2) quantify cross-layer spectral evolution as a token-wise transformation intensity; and (3) integrate this evolution signal to obtain a robust importance estimate. Unlike purely attention-driven or magnitude-based pruning strategies, our approach captures *representation dynamics* across layers. Furthermore, our framework is orthogonal to token merging methods and can be seamlessly combined with merging strategies such as ToMe [3]. This complementarity is particularly beneficial in video settings, where temporal redundancy amplifies token correlations. Extensive experiments on image and video benchmarks demonstrate that CLSE consistently improves the trade-off between efficiency and accuracy across multiple MLLMs. In summary, our contributions are threefold:

- We introduce the cross-layer spectral evolution framework, dubbed CLSE, as a token importance estimation principle for MLLMs, which characterizes how visual representations evolve from structural details to semantic abstractions across LLM layers.
- We design a training-free spectral-guided token pruning framework that operationalizes CLSE to identify and remove redundant tokens during inference while preserving tokens undergoing meaningful representational evolution.
- We conduct extensive experiments on both image- and video-based multimodal benchmarks, showing that CLSE achieves a superior trade-off between efficiency and accuracy under aggressive token reduction. Across multiple MLLMs, our method consistently reduces FLOPs, KV cache memory, and latency while maintaining competitive or improved performance.

2 Related Work

Training-Free Token Reduction for MLLMs. Multimodal Large Language Models (MLLMs) [7, 28, 30, 53] incur substantial overhead due to long visual token sequences. Reducing the number of visual tokens during inference has

emerged as an effective strategy for accelerating MLLMs without retraining. Existing training-free methods typically estimate token importance using heuristics derived from intermediate representations and can be broadly categorized into attention-based, similarity-based, diversity-based, and hybrid approaches. *Attention-based methods* estimate token importance directly from cross-modal attention between text and visual tokens. FastV [5] is a representative plug-and-play approach that ranks visual tokens using attention scores at a selected transformer layer and prunes tokens with the lowest attention magnitude for subsequent layers. Several follow-up works extend this idea using different attention aggregation or dynamic selection, strategies [6,14,16,26,31,41,42,45,50,56,62,65]. These approaches are computationally efficient and easy to integrate with existing MLLMs, but they generally rely on absolute attention magnitude as the ranking signal. *Similarity-based methods* identify redundant visual tokens by measuring similarity in feature space and removing tokens that provide overlapping information. For example, DART [49] detects redundant visual tokens through similarity-based matching and selectively retains a compact subset of representative tokens. Related approaches also employ clustering or feature distance criteria [18,21,39,54,58]. *Diversity-based methods* aim to maintain a representative set of visual tokens by maximizing feature diversity among the selected tokens. For instance, DivPrune [1] encourages diversity in the retained tokens to avoid redundancy and improve coverage of different visual regions. Other approaches formulate token selection as an entropy or diversity maximization problem to preserve informative structures in the visual representation [46, 61, 63]. While diversity criteria help prevent redundancy, they may not explicitly capture task relevance. *Hybrid methods* combine multiple criteria to improve token selection robustness. VisPruner [9] integrates attention-based importance with diversity-aware selection, while ToDre [24] jointly considers diversity and attention signals when selecting tokens. *Token merging approaches* reduce sequence length by merging visually similar tokens rather than removing them. Representative methods include LightVLM [17], CrossGet [38], VisionZip [55], and SparseVLM [64].A concurrent work V^2Drop [4] explores intrinsic cross-layer dynamics by progressively dropping 'lazy tokens' with minimal representational variation.

In contrast to existing methods that rely on instantaneous signals, we propose to measure *cross-layer spectral evolution* of tokens. By analyzing how token representations evolve across layers in the frequency domain, our approach captures structural-to-semantic transitions that are not visible from single-layer statistics, leading to more reliable identification of informative visual tokens.

Frequency Domain in Transformers. A growing body of work interprets Transformers through a frequency-domain lens. In language, FNet [23] replaces self-attention with Fourier mixing to accelerate computation, demonstrating that spectral operations can serve as efficient token mixers. VTC-LFC [47] compresses Vision Transformers by retaining only low-frequency spectral components of token representations. SFHformer [22] introduces frequency-domain token mixing via FFT to model global interactions with lower computational cost than stan-

dard self-attention, demonstrating the effectiveness of spectral operations in vision Transformers. In multimodal models, frequency-domain compression has also been explored; for instance, Fourier-VLM [43] compresses vision tokens by applying low-pass filtering in the frequency domain. Different from these approaches that apply spectral transforms for compression within a layer, we use cross-layer spectral evolution as an importance criterion, capturing whether a token undergoes meaningful detail-to-semantics transformation across layers.

3 Method

3.1 Preliminaries

Multimodal Large Language Models. An MLLM consists of a vision encoder f_v, a LLM decoder f_l, and a projection module $\phi(\cdot)$ that maps visual embeddings into the LLM token space. Given an image or video frames I and text prompt T, we obtain:

$$\mathbf{X}^0 = \phi(f_v(I)) \in \mathbb{R}^{N\times d}, \qquad \mathbf{Z}^0 = \mathrm{Embed}(T) \in \mathbb{R}^{M\times d}, \tag{1}$$

where N and M are the numbers of visual and text tokens and d is the shared embedding dimension. We concatenate tokens into a joint sequence processed by L Transformer layers:

$$\mathbf{H}^0 = [\mathbf{Z}^0; \mathbf{X}^0] \in \mathbb{R}^{(M+N)\times d}, \qquad \mathbf{H}^\ell = \mathrm{Transformer}^\ell(\mathbf{H}^{\ell-1}), \quad \ell = 1, \ldots, L. \tag{2}$$

For convenience, we denote the *visual slice* of layer ℓ as $\mathbf{X}^\ell \in \mathbb{R}^{N\times d}$.

Token Redundancy and Reduction. Visual tokens typically exhibit strong spatial and semantic redundancy. Not all tokens contribute equally to cross-modal reasoning: background regions or visually repetitive patches may have limited influence on the final output. Token reduction aims to select a subset of $K \ll N$ informative tokens:

$$\tilde{\mathbf{X}} = \mathcal{S}(\mathbf{X}), \quad \tilde{\mathbf{X}} \in \mathbb{R}^{K\times d}, \tag{3}$$

where $\mathcal{S}$ is a selection operator that preserves tokens with high semantic contribution while discarding redundant ones. After selection, the reduced multimodal sequence becomes:

$$\tilde{\mathbf{H}} = \left[\mathbf{Z}; \tilde{\mathbf{X}}\right], \tag{4}$$

which is then processed by subsequent Transformer layers.

Fast Fourier Transform (FFT). Let $\mathbf{X} \in \mathbb{R}^{H\times W}$ denote a 2D signal, where H and W are spatial height and width, and $X_{h,w}$ is the value at coordinate (h, w) with $h \in \{0, \ldots, H-1\}$ and $w \in \{0, \ldots, W-1\}$. The 2D Discrete Fourier Transform (DFT) maps $\mathbf{X}$ to a frequency representation $\hat{\mathbf{X}} \in \mathbb{C}^{H\times W}$:

$$\hat{X}_{u,v} = \sum_{h=0}^{H-1}\sum_{w=0}^{W-1} X_{h,w} e^{-j2\pi\left(\frac{uh}{H}+\frac{vw}{W}\right)}, \tag{5}$$

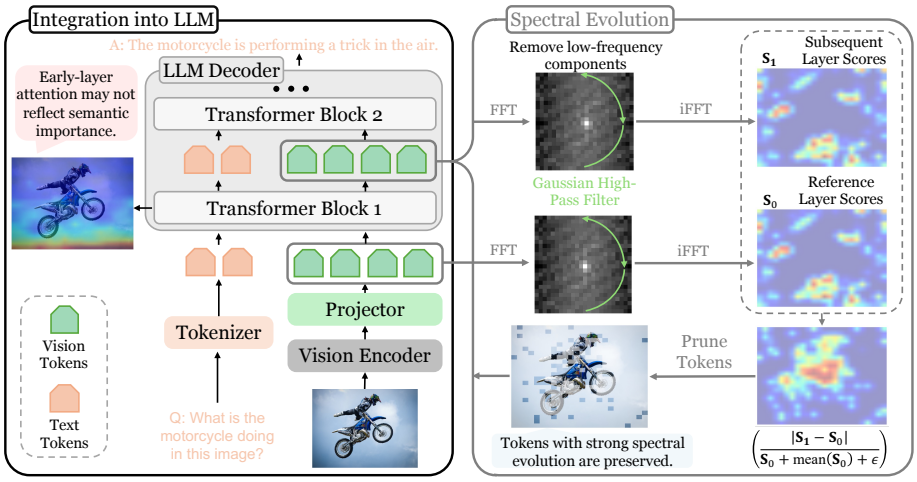


Fig. 2: Overview of the proposed spectral evolution-guided token pruning framework. **Left**: Cross-layer spectral evolution is computed by transforming layer-wise token representations into the frequency domain, suppressing low-frequency components via a Gaussian high-pass filter, and measuring the normalized spectral difference between adjacent layers. Tokens exhibiting stronger spectral redistribution are considered decision-critical and preserved. **Right**: The CLSE module is integrated into early LLM decoder layers to guide token selection. Compared with raw attention scores, cross-layer spectral evolution provides a more reliable signal for identifying semantically evolving tokens.

where (u, v) are frequency indices and $j = \sqrt{-1}$. We denote the forward and inverse transforms as $\hat{\mathbf{X}} = \mathcal{F}(\mathbf{X})$ and $\mathbf{X} = \mathcal{F}^{-1}(\hat{\mathbf{X}})$, respectively. The DFT decomposes $\mathbf{X}$ into orthogonal complex exponentials of increasing spatial frequency: low-frequency coefficients encode global, slowly varying structures, while high-frequency coefficients capture fine-grained spatial variations.

3.2 Cross-Layer Spectral Evolution (CLSE)

In Fig. 2, we present an intuitive overview of our CLSE. Let $N = H \cdot W$ and reshape the visual tokens $\mathbf{X}^{\ell} \in \mathbb{R}^{N \times d}$ into spatial feature maps $\mathbf{X}^{\ell}_{\text{2D}} \in \mathbb{R}^{H \times W \times d}$, we measure high-frequency structural energy instead of raw embedding magnitude. We first transform to the frequency domain and centre the spectrum:

$$\widehat{\mathbf{X}}^{\ell} = \text{fftshift}(\mathcal{F}(\mathbf{X}^{\ell}_{\text{2D}})). \tag{6}$$

To suppress low-frequency global patterns and highlight local structure, we apply a Gaussian high-pass mask $\mathbf{M} \in \mathbb{R}^{H \times W}$:

$$\mathbf{M}_{u,v} = 1 - \exp\left(-\frac{(u - c_u)^2 + (v - c_v)^2}{2(r \cdot c_u)^2}\right), \tag{7}$$

where (c_u, c_v) is the spectral centre and r is the cutoff ratio. We then invert the transform to obtain a spatial residual map and average over channels:

$$\mathbf{R}_\ell = \left| \mathcal{F}^{-1}\Big(\text{ifftshift}(\widehat{\mathbf{X}}^\ell \odot \mathbf{M}) \Big) \right|, \qquad \mathbf{S}_\ell = \text{Avg}_d(\mathbf{R}_\ell) \in \mathbb{R}^{H \times W}. \tag{8}$$

Flattening $\mathbf{S}_\ell$ over spatial positions yields $\mathbf{S}_\ell \in \mathbb{R}^N$, a per-token distribution of high-frequency energy. Crucially, $\mathbf{S}_\ell$ is *not* the full feature norm. Instead, it isolates band-limited residual energy after low-frequency suppression, where an ablation study evidenced this in Fig. 5(c).

CLSE measures how a token's band-limited structure changes across depth throughout the process of text-vision interaction in the transformer blocks. Background tokens tend to exhibit near-static residual patterns, while semantically engaged tokens often show pronounced redistribution. For a reference layer ℓ and a subsequent layer $\ell{+}1$, we define the evolution intensity:

$$\mathcal{I}^{(\ell)} = \text{clip}\left(\frac{|\mathbf{S}_{\ell+1} - \mathbf{S}_\ell|}{\mathbf{S}_\ell + \overline{\mathbf{S}}_\ell + \epsilon},\ 0,\ \tau_{\max} \right), \qquad \overline{\mathbf{S}}_\ell = \text{mean}(\mathbf{S}_\ell \text{ over tokens}), \tag{9}$$

where $\mathcal{I}^{(\ell)}$ is a relative metric to achieve scale-invariance. Dividing by $\mathbf{S}_\ell$ prevents tokens with large embedding magnitudes from dominating the signal, while the global mean $\overline{\mathbf{S}}_\ell$ acts as a regularizer to further suppress minor spectral fluctuations in inactive or redundant regions (where local energy is below the image-wide average). ϵ and $\tau_{\max}$ ensure numerical stability and outlier robustness.

To maximize prefilling speedups, we perform early-stage pruning in the LLM decoder. In our default setting, we compute CLSE from layer $0 \to 1$, which is between input embeddings $\mathbf{X}^0$ and the first decoder layer output $\mathbf{X}^1$, and execute pruning before layer 2. For each sample, we select the top-K tokens based on the evolution intensities:

$$\mathcal{K} = \text{TopK}(\mathcal{I}^{(\ell)}, K), \qquad \tilde{\mathbf{X}}^\ell = \mathbf{X}^\ell[\mathcal{K}] \in \mathbb{R}^{K \times d}, \tag{10}$$

and continue decoding with the truncated sequence $\tilde{\mathbf{H}}^\ell = [\mathbf{Z}^\ell; \tilde{\mathbf{X}}^\ell]$ for all subsequent layers. This yields a training-free, plug-and-play pruning module that significantly reduces computational overhead.

3.3 Theoretical Analysis

The above pruning rule admits a simple fidelity interpretation. Specifically, the calibrated spectral evolution score $\mathcal{I}^{(\ell)}$ quantifies the amount of cross-layer band-limited structural redistribution associated with each token. For any retained token set Ω with $|\Omega| = K$, let $P_\Omega(\mathbf{X}^\ell)$ denote the pruned token sequence. Then the pruning-induced perturbation of the subsequent decoder $g(\cdot)$ admits the bound:

$$\|g(\mathbf{X}^\ell) - g(P_\Omega(\mathbf{X}^\ell))\|_2 \leq \alpha \sum_{i \notin \Omega} \mathcal{I}_i^{(\ell)}, \tag{11}$$

where the right-hand side corresponds to the discarded spectral evolution mass. Therefore, selecting the top-K tokens by CLSE minimizes this upper bound

Table 1: Performance comparison with baselines on nine image understanding benchmarks with LLaVA-1.5-7B. The vanilla 576 tokens are pruned to 192, 128, and 64.

Method	GQA	MMB	MMB_{CN}	MME	POPE	SQA	VQA_{Text}	VizWiz	OCR_{Bench}	Avg.
LLaVA-1.5-7B	*Upper Bound, 576 Tokens* **(100%)**									
Vanilla	61.9	64.7	58.1	1862	85.9	69.5	58.2	50.0	29.7	100.0%
	Retain 192 Tokens (↓ **66.7%**)									
ToMe (ICLR23)	54.3	60.5	53.2	1563	65.2	68.0	52.1	50.9	25.8	84.8%
FastV (ECCV24)	52.7	61.2	57.0	1612	67.3	67.1	52.5	50.8	29.1	92.1%
PruMerge (ICCV25)	54.3	59.6	52.9	1632	71.3	67.9	54.3	50.1	25.3	90.8%
PDrop (CVPR25)	57.1	63.2	56.8	1766	82.3	68.8	56.1	**51.1**	28.7	96.9%
HiRED (AAAI25)	58.7	62.8	54.7	1737	82.8	68.4	47.4	50.1	19.0	91.0%
SparseVLM (ICML25)	57.6	62.5	53.7	1721	83.6	69.1	56.1	50.5	29.2	96.3%
FiCoCo-V (AAAI26)	58.5	62.3	55.3	1732	82.5	67.8	55.7	51.0	28.6	96.2%
CLSE (Ours)	**61.5**	**63.6**	**57.2**	**1817**	**85.1**	**70.1**	**57.5**	50.6	**30.1**	**99.4%**
	Retain 128 Tokens (↓ **77.8%**)									
ToMe (ICLR23)	52.4	53.3	50.7	1343	62.8	59.6	49.1	**51.6**	25.0	84.1%
FastV (ECCV24)	49.6	56.1	56.4	1490	59.6	60.2	50.6	51.3	28.5	87.2%
PruMerge (ICCV25)	53.3	58.1	51.7	1554	67.2	67.1	54.3	50.3	24.8	88.9%
PDrop (CVPR25)	56.0	61.1	**56.6**	1644	82.3	68.3	55.1	51.0	**28.7**	95.3%
HiRED (AAAI25)	57.2	61.5	53.6	1710	79.8	68.1	46.1	51.3	19.1	89.8%
SparseVLM(ICML25)	56.0	60.0	51.1	1696	80.5	67.1	54.9	51.4	28.0	93.7%
FiCoCo-V (AAAI26)	57.6	61.1	54.3	1711	82.2	68.3	55.6	49.4	26.2	94.3%
CLSE (Ours)	**60.9**	**63.0**	55.8	**1816**	**84.4**	**70.1**	**56.6**	51.0	28.4	**98.1%**
	Retain 64 Tokens (↓ **88.9%**)									
ToMe (ICLR23)	48.6	43.7	47.1	1138	52.5	50.0	45.3	**52.6**	22.7	67.4%
FastV (ECCV24)	46.1	48.0	52.7	1256	48.0	51.1	47.8	50.8	24.5	78.0%
PruMerge (ICCV25)	51.9	55.3	49.1	1549	65.3	68.1	54.0	50.1	25.0	87.5%
PDrop (CVPR25)	41.9	33.3	50.5	1092	55.9	68.6	45.9	50.7	25.0	77.0%
HiRED (AAAI25)	54.6	60.2	51.4	1599	73.6	68.2	44.2	50.2	19.1	86.6%
SparseVLM (ICML25)	52.7	56.2	46.1	1505	75.1	62.2	51.8	50.1	18.0	84.3%
FiCoCo-V (AAAI26)	52.4	60.3	53.0	1591	76.0	68.1	53.6	49.8	22.6	89.8%
CLSE (Ours)	**58.3**	**61.9**	**54.3**	**1809**	**78.3**	**69.6**	**55.5**	50.9	**25.1**	**94.8%**

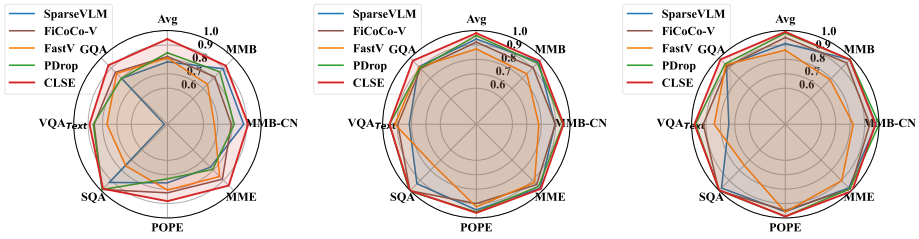
among all subsets of size K, *i.e.,* it preserves the maximum amount of cross-layer spectral evolution under a fixed token budget. A derivation is provided in the appendix.

4 Experiments

To validate the effectiveness of our CLSE framework, we conduct experiments on image and video question-answering datasets with different MLLM variants.

4.1 Image VQA Tasks

Experiment setup. We evaluate CLSE on nine diverse multimodal benchmarks spanning visual reasoning (GQA [19], SQA [34], VizWiz [13]), hallucination detection (POPE [25]), OCR (VQA_{TEXT} [40], OCR_{Bench} [33]), and holistic understanding (MME [11], MMB [32], MMB_{CN} [32], MMVet [59]). We compare against seven state-of-the-art training-free methods: ToMe [3], FastV [5], PDrop [51], SparseVLM [64], FiCoCo [15], HiRED [2], and PruMerge [37]. To



(a) 64 Tokens **(b)** 128 Tokens **(c)** 192 Tokens

Fig. 3: Radar plots comparing pruning methods on LLaVA-1.5-13B under different token budgets. All metrics, including the overall average (Avg), are normalized to the full-token model. Across all configurations, CLSE achieves the most balanced performance and maintains the highest overall accuracy under aggressive compression.

assess cross-model generalizability, we evaluate on LLaVA-1.5-7B [30], LLaVA-1.5-13B [30], LLaVA-Next-7B [29], Qwen2-VL-7B [44], and Qwen2-VL-72B [44].

Implementation Details. We apply CLSE to each model's standard inference pipeline without modifying model weights or requiring additional training. Unlike other methods that perform pruning in later layers, for example, FastV [5] performs at layer 3, CLSE performs token pruning at the first layer of the LLM decoder. All experiments are conducted on NVIDIA RTX PRO 6000 GPU.

Main Results. Table 1 reports results on LLaVA-1.5-7B under three token-retention budgets (192, 128, and 64 tokens), corresponding to pruning ratios of 66.7%, 77.8%, and 88.9%, respectively. CLSE applies spectral evolution scoring with simple hard pruning and no token recycling. Despite this simplicity, CLSE achieves the best overall average of 99.4%, 98.1%, and 94.8% across the three budgets, consistently outperforming both hard-pruning methods (FastV, PDrop) and token-merging methods (SparseVLM, PruMerge, FiCoCo-V). This demonstrates that spectral evolution provides a more reliable token-importance criterion than attention- or magnitude-based strategies, regardless of whether the baseline uses hard pruning or token merging.

Scaling to Larger Models. We further conduct experiments on the larger LLaVA-1.5-13B under $K = 192, 128, 64$ tokens. Figure 3 shows that CLSE maintains the most balanced performance profile and achieves the strongest overall average across all budgets. The advantage becomes more evident as compression increases, where competing methods exhibit larger task-specific degradation. These results demonstrate that spectral evolution provides stable token importance estimation and scales effectively to larger models under aggressive pruning. Additionally, we provide an experimental comparison with baseline methods on an even larger model Qwen2-VL-72B in the appendix due to the page limit.

Generalization to Qwen2-VL-7B. Table 2 shows CLSE can generalize well to Qwen2-VL-7B under three token ratios. CLSE achieves 97.6%, 95.8% and 90.5% average accuracy relative to the vanilla model on three ratios. Under 66.7%

Table 2: Performance comparison with baseline methods under different token configurations on Qwen2-VL-7B.

Method	GQA	MMB	MMB_{CN}	MME	POPE	SQA	VQA_{Text}	Avg.
Qwen2-VL-7B	*Upper Bound, All Tokens* **(100%)**							
Vanilla	62.2	79.1	77.8	2296	87.7	85.4	81.4	100%
Qwen2-VL-7B	*Token Reduction* (↓ **66.7%**)							
+ FastV (ECCV24)	58.6	74.5	74.3	2083	85.6	83.1	76.1	94.9%
+ PDrop(CVPR25)	58.5	72.3	71.2	1991	86.1	82.8	76.4	93.1%
+ HiRED (AAAI25)	59.9	73.8	73.4	2158	86.6	80.7	71.0	94.1%
+ SparseVLM(ICML25)	59.2	**77.1**	**76.0**	2181	85.9	83.5	77.8	96.7%
+ CLSE (ours)	**61.0**	76.5	75.5	**2254**	**86.7**	**84.2**	**78.3**	**97.6%**
Qwen2-VL-7B	*Token Reduction* (↓ **77.8%**)							
+ FastV (ECCV24)	56.2	70.1	72.4	1957	82.7	82.3	73.0	91.3%
+ PDrop(CVPR25)	56.4	67.5	67.2	1945	84.2	82.0	73.7	89.9%
+ HiRED (AAAI25)	58.3	70.5	70.7	2086	85.4	79.6	65.4	90.8%
+ SparseVLM(ICML25)	54.9	73.2	73.6	2047	81.1	83.2	72.2	91.9%
+ CLSE (ours)	**59.8**	**74.1**	**74.2**	**2209**	**85.5**	**83.5**	**76.6**	**95.8%**
Qwen2-VL-7B	*Token Reduction* (↓ **88.9%**)							
+ FastV (ECCV24)	50.6	61.8	62.5	1760	72.2	80.5	68.6	82.2%
+ PDrop(CVPR25)	52.5	59.2	58.3	1756	74.1	80.2	67.7	81.5%
+ HiRED (AAAI25)	55.5	64.3	64.3	1834	**82.8**	75.7	56.5	83.6%
+ SparseVLM(ICML25)	47.5	59.5	61.1	1624	68.0	80.7	61.7	78.4%
+ CLSE (ours)	**56.1**	**69.3**	**69.2**	**2008**	81.2	**82.6**	**73.4**	**90.5%**

Table 3: Comparative experiments are performed on LLaVA-Next-7B.

Method	GQA	MMB	MMB-CN	MME	POPE	SQA	VQA^{Text}	VizWiz	OCRBench	Avg.
LLaVA-Next-7B	*Upper Bound, 2880 Tokens* **(100%)**									
Vanilla	64.2	67.4	60.6	1851	86.5	70.1	64.9	57.6	51.7	100.0%
LLaVA-Next-7B	*Retain 320 Tokens* (↓ **88.9%**)									
FastV (ECCV24)	55.9	61.6	51.9	1661	71.7	62.8	55.7	53.1	37.4	86.3%
HiRED (AAAI25)	59.3	64.2	55.9	1690	83.3	66.7	58.8	54.2	40.4	91.7%
PruMerge (ICCV2025)	53.6	61.3	55.3	1534	60.8	66.4	50.6	54.0	14.6	79.2%
SparseVLM (ICML25)	56.1	60.6	54.5	1533	82.4	66.1	58.4	52.0	27.0	85.8%
PDrop (CVPR25)	56.4	63.4	56.2	1663	77.6	67.5	54.4	54.1	25.9	86.5%
CLSE (Ours)	**62.6**	**67.1**	**58.3**	**1794**	**84.4**	**67.9**	**58.9**	**56.1**	**41.4**	**94.7%**

pruning, CLSE achieves the highest average performance (97.6%), outperforming FastV and PDrop across most benchmarks, with strong results on GQA (61.0), MMB (76.5), POPE (86.7), and SQA (84.2). As compression increases to 77.8%, CLSE maintains the same high average score (95.8%), demonstrating stable performance despite further token reduction. Even under extreme pruning (88.9%), CLSE continues to outperform competing methods with an average of 90.5%, showing clear advantages on reasoning-heavy tasks such as GQA, POPE, and VQA_{Text}. Across all settings, CLSE consistently delivers the best overall performance and exhibits strong robustness as token budgets shrink, validating that cross-layer spectral evolution generalizes beyond LLaVA and remains effective across different MLLMs architectures.

Generalization to LLaVA-Next-7B. To further evaluate the generalization ability of our method, we conduct comparative experiments on LLaVA-Next-7B. The results are summarized in Table 3. The vanilla model processes 2880 visual tokens and serves as the upper-bound performance reference. Under a strict token budget of 320 tokens (88.9% token reduction), we compare CLSE with several recent token reduction methods including FastV, HiRED, PruMerge, SparseVLM, and PDrop. Overall, CLSE achieves the best performance among all competing methods, reaching 94.7% of the vanilla performance, significantly outperforming other pruning approaches. In particular, CLSE preserves strong performance on reasoning-intensive benchmarks such as GQA (62.6) and MMB (67.1), and maintains competitive results across multimodal evaluation suites including MMB-CN, MME, POPE, and SQA. Compared with other pruning methods that suffer noticeable degradation under aggressive token reduction, CLSE consistently maintains higher accuracy across most benchmarks. Notably, CLSE achieves these improvements while operating under the same token budget as other methods, demonstrating that cross-layer spectral evolution provides a more reliable criterion for identifying informative visual tokens. This result indicates that modeling representation dynamics across layers, rather than relying on instantaneous signals, leads to more effective token selection in multimodal large language models.

4.2 Video Understanding Tasks

Experimental Settings. We apply CLSE to Video-LLaVA [27] and evaluate its performance on four representative video question-answering benchmarks: MSVD-QA [52], MSRVTT-QA [52], TGIF-QA [20], and ActivityNet-QA [60]. We compare against several advanced baselines: FastV [5], which performs hard pruning; PDrop [51], which adopts a pyramid-shaped progressive pruning schedule; SparseVLM [64] and FiCoCo-V [15], which execute token merging. We report Accuracy (%) and Score (0–5) using GPT-4o-mini and Gemini-2.5-Flash as assistant evaluators.

Main Results. Table 4 reports results when compressing video tokens from 2048 to 194 (over 90% reduction). CLSE applies per-frame 2D FFT to compute spectral evolution scores and performs hard pruning, as in FastV. Despite this simplicity, spectral evolution provides a substantially better importance signal: CLSE achieves 35.7% average accuracy, compared to 26.1% for FastV and 27.3% for PDrop, demonstrating that per-frame spatial frequency dynamics capture informative visual content more reliably than attention-based selection. At such extreme compression ratios, however, hard pruning inevitably discards temporal content that is difficult to recover. We therefore introduce CLSE-M, which recalls the top-scored pruned tokens by their spectral evolution scores and compresses them into compact representatives via density-peak clustering. CLSE-M achieves 38.7% average accuracy (GPT), matching the vanilla model (38.7%), and outperforms SparseVLM—which adopts the same token recycling strategy but relies on attention-based scoring—by 0.6 points in accuracy (38.7% vs. 38.1%), con-

Table 4: Performance comparison on video understanding benchmarks with Video-LLaVA-7B. We consistently prune video tokens from 2048 to 194 (over 90% reduction). CLSE-M denotes CLSE with token merging, where pruned tokens are recalled and clustered into compact representatives based on their spectral evolution scores. GPT-4o-mini (GPT) and Gemini-2.5-Flash (Gem) are adopted as assistant evaluators.

Method	Eval	TGIF		MSVD		MSRVTT		ActivityNet		Avg.	
		Acc.	Score	Acc.	Score	Acc.	Score	Acc.	Score	Acc.	Score
Video-LLaVA-7B	GPT	11.3	1.19	58.9	3.31	42.7	2.69	42.0	2.20	38.7	2.35
	Gem	14.2	0.77	57.3	2.93	47.5	2.32	42.5	2.12	40.4	2.04
Hard Pruning											
FastV	GPT	2.8	1.16	35.9	2.36	31.4	2.19	34.2	1.83	26.1	1.88
	Gem	2.3	0.14	35.5	1.88	34.3	1.69	34.2	1.74	26.6	1.36
PDrop	GPT	5.0	0.96	42.2	2.59	27.9	1.95	34.3	1.78	27.3	1.82
	Gem	6.1	0.37	52.2	2.54	34.6	1.65	39.4	1.93	33.0	1.62
FiCoCo-V	GPT	5.6	1.18	42.1	2.68	37.6	2.47	**43.1**	2.23	31.9	2.13
	Gem	6.7	0.38	51.2	2.51	44.3	2.10	**44.9**	2.20	36.7	1.79
CLSE	GPT	**9.6**	1.06	**54.1**	3.11	**39.3**	2.56	39.9	2.03	**35.7**	2.19
	Gem	**12.9**	0.69	**56.3**	2.88	**47.7**	2.27	39.9	1.99	**39.2**	1.95
w/ Token Merging											
SparseVLM	GPT	10.7	1.03	56.0	3.20	41.9	2.63	43.6	2.21	38.1	2.27
	Gem	13.1	0.71	58.2	3.05	48.0	2.38	44.0	2.20	40.8	2.09
CLSE-M	GPT	**11.2**	1.05	**57.0**	3.25	**42.2**	2.61	**44.6**	2.25	**38.7**	2.29
	Gem	**14.7**	0.79	**60.2**	3.07	**49.5**	2.35	**45.0**	2.24	**42.3**	2.11

firming that spectral evolution provides a more reliable signal for identifying critical visual dynamics.

4.3 Analysis

Efficiency Analysis. Table 5 compares inference efficiency on LLaVA-1.5-7B (192 tokens) and Video-LLaVA-7B (194 tokens). On images, CLSE achieves the lowest FLOPs (**43.9%**) and smallest KV cache (**56.0 MB**) with the highest performance (**99.6%**) at 1.50× speedup; FastV is faster (1.69×) but drops to 92.1%. On video, CLSE-M matches SparseVLM in FLOPs (**10.7%**) and KV cache (**121.1 MB**) while achieving perfect performance retention (**100%** vs. 98.4%) at 2.73× speedup; FastV again leads in raw speed (3.32×) but suffers severe performance drop.

Efficiency Compared with Attention Scoring. In Fig. 4, we analyze the computational efficiency of our FFT-based scoring relative to the attention-extraction overhead Δ, benchmarked on a single simulated attention layer using LLaMA-3-8B's configuration [12]. In modern inference engines, optimized kernels

Table 5: Efficiency–performance trade-off comparison on LLaVA-1.5-7B and Video-LLaVA-7B under comparable token budgets. CLSE reduces FLOPs and KV cache more aggressively while preserving accuracy, achieving the most balanced prefilling acceleration among all methods.

Methods	Tokens ↓	Prefilling (MS) ↓	FLOPs ↓	KV Cache (MB) ↓	Perform. ↑	Throughput↑ (samples/s)	Speedup ↑ (Prefilling)
LLaVA-1.5-7B	576	41.9	100%	313.0	100%	19.4	1.00×
+ FastV	192	**24.8**	46.9%	121.0	92.1%	**30.3**	**1.69**×
+ PDrop	192	31.9	74.0%	211.2	96.9%	24.6	1.31×
+ SparseVLM	192	33.0	49.7%	128.2	96.3%	23.7	1.27×
+ **CLSE**(ours)	192	27.8	**43.9%**	**56.0**	**99.4%**	27.9	1.50×
Video-LLaVA-7B	2048	114.0	100%	1053	100%	7.6	1.00×
+ FastV	194	**34.3**	20.6%	212.9	67.4%	**19.4**	**3.32**×
+ PDrop	194	45.9	20.6%	209.2	70.5%	16.0	2.48×
+ SparseVLM	194	40.0	**10.7%**	**121.1**	98.4%	18.1	2.85×
+ **CLSE-M**(ours)	194	41.7	**10.7%**	**121.1**	**100%**	17.3	2.73×

such as Scaled Dot-Product Attention (SDPA) achieve high throughput by avoiding the materialization of the full $N \times N$ attention matrix. Consequently, any pruning strategy relying on dense attention weights necessitates falling back to a sub-optimal "Eager" implementation, which reintroduces a quadratic latency penalty $O(N^2)$. As illustrated, while the extraction overhead Δ (defined as $\text{Latency}_{\text{Eager}} - \text{Latency}_{\text{SDPA}}$) escalates quadratically with sequence length N, our CLSE module maintains a near-linear $O(N \log N)$ complexity due to the efficiency of the Fast Fourier Transform. Even at a spatial resolution of 64×64 ($N = 4096$), the latency of our spectral scoring remains significantly below the Δ threshold. This gap demonstrates that spectral evolution also offers a more hardware-efficient pathway for token reduction.

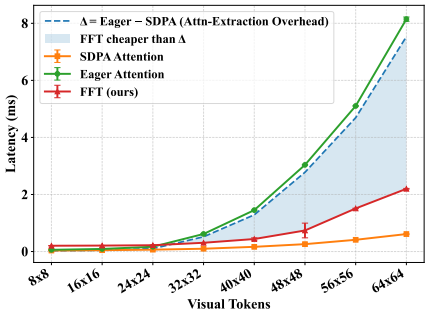
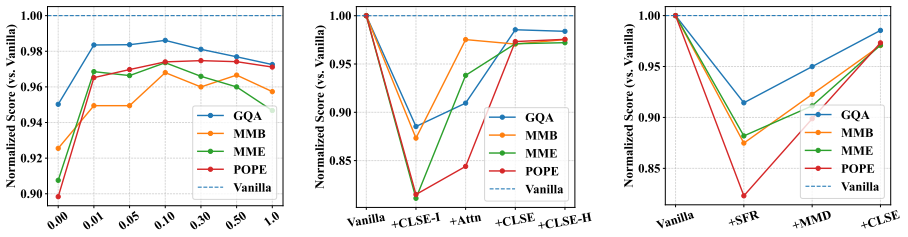


Fig. 4: Latency Comparison to Attention Scoring.

Ablation Studies. Fig. 5 presents comprehensive ablation analyses on LLaVA-1.5-7B with 192 retained tokens. All results are normalized to the full-token model, where 1.0 indicates matching full performance. Fig. 5(a) evaluates the impact of the Gaussian high-pass cutoff parameter. In all experimental scenarios, we set the cutoff to 0.1. Performance improves substantially when moving from $r = 0.00$ to small values, indicating that mild suppression of low-frequency components effectively reduces redundancy. Across benchmarks, moderate cutoff values consistently yield the strong CLSE results. Fig. 5(b) evaluates the core components of our framework. We observe a catastrophic performance collapse in CLSE-Inertia (CLSE-I), which retains tokens with the *lowest* spectral evolution scores (the most static tokens), serving as counterfactual evidence that depth-wise spectral evolution is the primary driver of semantic activity.



(a) Cutoff Ratio r **(b)** Spectral Evolution **(c)** Token Importance Metrics

Fig. 5: Ablation studies on LLaVA-1.5-7B pruned after the first layer of LLM decoder with 192 retained tokens. Scores are normalized to the Vanilla model. (a) Effect of the Gaussian high-pass cutoff ratio. (b) Effect of spectral evolution and its integration with attention. (c) Impact of the magnitude measures for cross-layer component analysis.

In contrast, while pure attention-based pruning suffers from positional bias, our CLSE criterion significantly stabilizes performance. To further fortify the system, CLSE-Hybrid (CLSE-H) integrates CLSE with attention weights that can produce on-par performance like CLSE. Fig. 5(c) evaluates CLSE against alternative depth-wise metrics. We compare with Spatial Feature Residual (SFR), which measures the L1-norm of feature residuals, and Mean Magnitude Difference (MMD), which monitors changes in channel-wise mean activations. SFR exhibits performance collapse as its reliance on raw pixel-level intensity makes it susceptible to high-frequency artifacts. MMD, while slightly more stable, remains a coarse-grained heuristic. By collapsing high-dimensional token dynamics into a scalar magnitude, it fails to capture the structural redistribution occurring within the frequency spectrum. In contrast, CLSE consistently establishes a new performance upper bound.

Qualitative Analysis. Fig. 6 visualizes token importance maps under different strategies. Attention-based scores tend to spread over large background regions and occasionally over-emphasize visually salient but semantically irrelevant areas. In contrast, CLSE concentrates on structurally meaningful regions that actively evolve across layers, such as key objects or human subjects. When the cutoff ratio is set to $r = 0.1$, the maps exhibit sharper localization and reduced background noise compared to $r = 0$, confirming that moderate suppression of low-frequency components enhances semantic discrimination. These visualizations support our quantitative findings that cross-layer spectral evolution provides a more reliable token importance signal than raw attention.

Additional Experiments in Appendix. We provide complementary analyses in the appendix to further validate CLSE. These include experimental setup details (*e.g.,* baseline methods, implementation, and datasets), details of a CLSE variant with token merging, ablation on pruning at different layers, experiments on additional MLLMs (InternVL, Qwen2-VL-72B), complete numeric results for

Fig. 6: Qualitative comparison of token importance maps. Top row: original images. Second row: attention-based token scores. Third and fourth rows: CLSE without frequency suppression ($r = 0$) and with moderate suppression ($r = 0.1$), respectively. Compared to attention, CLSE highlights structurally evolving and semantically relevant regions while suppressing redundant background areas. Moderate frequency filtering further improves focus on task-relevant objects.

LLaVA-1.5-13B, a detailed computational complexity analysis, channel dimension impact on run time, and impact of 3D vs 2D FFT on video MLLMs.

5 Conclusion

In this work, we introduced a token pruning framework, namely Cross-Layer Spectral Evolution (CLSE), that models how token representations evolve across depth, capturing the structural transition from high-frequency detail to low-frequency semantic abstraction. Extensive experiments on image and video benchmarks demonstrate that CLSE consistently achieves superior accuracy–efficiency trade-offs across multiple architectures and compression levels. In particular, the performance gap widens under aggressive pruning, highlighting the robustness of spectral evolution as a token importance signal. We hope this work encourages further exploration of spectral representations for efficient multimodal reasoning.

Acknowledgements

This work is supported by the National Natural Science Foundation of China under Grant No. 62502429, and the Zhejiang Key Laboratory Project (2024E10001)

References

1. Alvar, S.R., Singh, G., Akbari, M., Zhang, Y.: Divprune: Diversity-based visual token pruning for large multimodal models. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 9392–9401 (2025)
2. Arif, K.H.I., Yoon, J., Nikolopoulos, D.S., Vandierendonck, H., John, D., Ji, B.: Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 1773–1781 (2025)
3. Bolya, D., Fu, C.Y., Dai, X., Zhang, P., Feichtenhofer, C., Hoffman, J.: Token merging: Your vit but faster. In: The Eleventh International Conference on Learning Representations (2023)
4. Chen, J., Liu, X., Wen, Z., Wang, Y., Huang, S., Chen, H.: Variation-aware vision token dropping for faster large vision-language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3489–3499 (2026)
5. Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., Chang, B.: An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In: European Conference on Computer Vision. pp. 19–35. Springer (2024)
6. Chen, Z., Cai, Y., Guo, J., Cai, T., Yin, J., Chen, Z.: Accelerating multimodal large language models with prior-corrected token reduction. In: European Conference on Computer Vision. Springer (2026)
7. Chen, Z., Wu, J., Wang, W., Su, W., Chen, G., Xing, S., Zhong, M., Zhang, Q., Zhu, X., Lu, L., et al.: 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. pp. 24185–24198 (2024)
8. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
9. Fan, Y., Zhao, A., Fu, J., Tong, J., Su, H., Pan, Y., Zhang, W., Shen, X.: Visipruner: Decoding discontinuous cross-modal dynamics for efficient multimodal llms. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 18896–18913 (2025)
10. Fayyaz, M., Koohpayegani, S.A., Jafari, F.R., Sengupta, S., Joze, H.R.V., Sommerlade, E., Pirsiavash, H., Gall, J.: Adaptive token sampling for efficient vision transformers. In: European conference on computer vision. pp. 396–414. Springer (2022)
11. Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al.: Mme: A comprehensive evaluation benchmark for multimodal large language models. In: The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track (2025)
12. Grattafiori, A., Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., Mathur, A., Schelten, A., Vaughan, A., et al.: The llama 3 herd of models. arXiv preprint arXiv:2407.21783 (2024)
13. Gurari, D., Li, Q., Stangl, A.J., Guo, A., Lin, C., Grauman, K., Luo, J., Bigham, J.P.: Vizwiz grand challenge: Answering visual questions from blind people. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3608–3617 (2018)

14. Han, Y., Liu, X., Ding, P., Wang, D., Chen, H., Yan, Q., Huang, S.: Rethinking token reduction in mllms: Towards a unified paradigm for training-free acceleration. arXiv preprint arXiv:2411.17686 **2**(3) (2024)
15. Han, Y., Liu, X., Zhang, Z., Ding, P., Chen, J., Wang, D., Chen, H., Yan, Q., Huang, S.: Filter, correlate, compress: Training-free token reduction for mllm acceleration. In: Proceedings of the 40th AAAI Conference on Artificial Intelligence (2025)
16. He, Y., Chen, F., Liu, J., Shao, W., Zhou, H., Zhang, K., Zhuang, B.: Zipvl: Efficient large vision-language models with dynamic token sparsification. arXiv preprint arXiv:2410.08584 (2024)
17. Hu, L., Shang, F., Feng, W., Wan, L.: Lightvlm: Acceleraing large multimodal models with pyramid token merging and kv cache compression. arXiv preprint arXiv:2509.00419 (2025)
18. Huang, K., Zou, H., Xi, Y., Wang, B., Xie, Z., Yu, L.: Ivtp: Instruction-guided visual token pruning for large vision-language models. In: European conference on computer vision. pp. 214–230. Springer (2024)
19. Hudson, D.A., Manning, C.D.: 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. pp. 6700–6709 (2019)
20. Jang, Y., Song, Y., Yu, Y., Kim, Y., Kim, G.: Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 2758–2766 (2017)
21. Jeddi, A., Baghbanzadeh, N., Dolatabadi, E., Taati, B.: Similarity-aware token pruning: Your vlm but faster. arXiv preprint arXiv:2503.11549 (2025)
22. Jiang, X., Zhang, X., Gao, N., Deng, Y.: When fast fourier transform meets transformer for image restoration. In: European conference on computer vision. pp. 381–402. Springer (2024)
23. Lee-Thorp, J., Ainslie, J., Eckstein, I., Ontanon, S.: Fnet: Mixing tokens with fourier transforms. In: Proceedings of the 2022 Conference of the north American chapter of the Association for Computational Linguistics: human language technologies. pp. 4296–4313 (2022)
24. Li, D., Yang, Z., Lu, S.: Todre: Visual token pruning via diversity and task awareness for efficient large vision-language models. arXiv e-prints pp. arXiv–2505 (2025)
25. Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W.X., Wen, J.R.: Evaluating object hallucination in large vision-language models. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing. pp. 292–305 (2023)
26. Liang, X., Guan, C., Lu, J., Chen, H., Wang, H., Hu, H.: Dynamic token reduction during generation for vision language models. arXiv preprint arXiv:2501.14204 (2025)
27. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the 2024 conference on empirical methods in natural language processing. pp. 5971–5984 (2024)
28. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 26296–26306 (2024)
29. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge. In: Advances in Neural Information Processing Systems (NeurIPS) (2024)
30. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: Advances in Neural Information Processing Systems (NeurIPS) (2023)

31. Liu, Y., Wu, F., Li, R., Tang, Z., Li, K.: Par: Prompt-aware token reduction method for efficient large multimodal models. arXiv preprint arXiv:2410.07278 (2024)
32. Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al.: Mmbench: Is your multi-modal model an all-around player? In: European conference on computer vision. pp. 216–233. Springer (2024)
33. Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.C., Liu, C.L., Jin, L., Bai, X.: Ocrbench: on the hidden mystery of ocr in large multimodal models. Science China Information Sciences **67**(12), 220102 (2024)
34. Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A.: Learn to explain: Multimodal reasoning via thought chains for science question answering. Advances in Neural Information Processing Systems **35**, 2507–2521 (2022)
35. Nguyen, T., Pham, M., Nguyen, T., Nguyen, K., Osher, S., Ho, N.: Fourierformer: Transformer meets generalized fourier integral theorem. Advances in Neural Information Processing Systems **35**, 29319–29335 (2022)
36. Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., Hsieh, C.J.: Dynamicvit: Efficient vision transformers with dynamic token sparsification. Advances in neural information processing systems **34**, 13937–13949 (2021)
37. Shang, Y., Cai, M., Xu, B., Lee, Y.J., Yan, Y.: Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 22857–22867 (2025)
38. Shi, D., Tao, C., Rao, A., Yang, Z., Yuan, C., Wang, J.: Crossget: Cross-guided ensemble of tokens for accelerating vision-language transformers. arXiv preprint arXiv:2305.17455 (2023)
39. Si, G., Yin, H., Li, X., Liao, W., He, T., Peng, P., Zhu, W., et al.: Infoprune: Revisiting visual token pruning from an information-theoretic perspective
40. Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., Rohrbach, M.: Towards vqa models that can read. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 8317–8326 (2019)
41. Song, D., Wang, W., Chen, S., Wang, X., Guan, M.X., Wang, B.: Less is more: A simple yet effective token reduction method for efficient multi-modal llms. In: Proceedings of the 31st International Conference on Computational Linguistics. pp. 7614–7623 (2025)
42. Sun, Z., Ma, Y., Liu, G., Chen, Y., Tang, X., Hu, Y., Xu, Y.: Ivc-prune: Revealing the implicit visual coordinates in lvlms for vision token pruning. arXiv preprint arXiv:2602.03060 (2026)
43. Wang, H., Kai, J., Bai, H., Hou, L., Jiang, B., He, Z., Lin, Z.: Fourier-vlm: Compressing vision tokens in the frequency domain for large vision-language models. arXiv preprint arXiv:2508.06038 (2025)
44. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, Z., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
45. Wang, Q., Ye, H., Chung, M.Y., Liu, Y., Lin, Y., Kuo, M., Ma, M., Zhang, J., Chen, Y.: Corematching: A co-adaptive sparse inference framework with token and neuron pruning for comprehensive acceleration of vision-language models. arXiv preprint arXiv:2505.19235 (2025)
46. Wang, Y., Wu, J., Ni, Z., Yang, C., Liu, Y., Yang, L., Zhou, Y., Wen, Y., He, L.: Entropyprune: Matrix entropy guided visual token pruning for multimodal large language models. arXiv preprint arXiv:2602.17196 (2026)

47. Wang, Z., Luo, H., Wang, P., Ding, F., Wang, F., Li, H.: Vtc-lfc: Vision transformer compression with low-frequency components. Advances in Neural Information Processing Systems **35**, 13974–13988 (2022)
48. Wen, Z., Gao, Y., Li, W., He, C., Zhang, L.: Token pruning in multimodal large language models: Are we solving the right problem? In: Findings of the Association for Computational Linguistics: ACL 2025. pp. 15537–15549 (2025)
49. Wen, Z., Gao, Y., Wang, S., Zhang, J., Zhang, Q., Li, W., He, C., Zhang, L.: Stop looking for "important tokens" in multimodal language models: Duplication matters more. In: Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing. pp. 9972–9991 (2025)
50. Xing, L., Wang, A.J., Yan, R., Shu, X., Tang, J.: Vision-centric token compression in large language model. arXiv preprint arXiv:2502.00791 (2025)
51. Xing, L., Huang, Q., Dong, X., Lu, J., Zhang, P., Zang, Y., Cao, Y., He, C., Wang, J., Wu, F., et al.: Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. In: In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (2025)
52. Xu, D., Zhao, Z., Xiao, J., Wu, F., Zhang, H., He, X., Zhuang, Y.: Video question answering via gradually refined attention over appearance and motion. In: Proceedings of the 25th ACM international conference on Multimedia. pp. 1645–1653 (2017)
53. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
54. Yang, C., Sui, Y., Xiao, J., Huang, L., Gong, Y., Li, C., Yan, J., Bai, Y., Sadayappan, P., Hu, X., et al.: Topv: Compatible token pruning with inference time optimization for fast and low-memory multimodal vision language model. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 19803–19813 (2025)
55. Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., Jia, J.: Visionzip: Longer is better but not necessary in vision language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19792–19802 (2025)
56. Ye, W., Wu, Q., Lin, W., Zhou, Y.: Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 39, pp. 22128–22136 (2025)
57. Yin, H., Vahdat, A., Alvarez, J.M., Mallya, A., Kautz, J., Molchanov, P.: A-vit: Adaptive tokens for efficient vision transformer. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 10809–10818 (2022)
58. Yu, H., Li, W., Qu, X., Wang, S., Chen, J., Zhu, J.: Visiontrim: Unified vision token compression for training-free mllm acceleration. arXiv preprint arXiv:2601.22674 (2026)
59. Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., Wang, L.: Mm-vet: Evaluating large multimodal models for integrated capabilities. In: Forty-first International Conference on Machine Learning (2024)
60. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. vol. 33, pp. 9127–9134 (2019)
61. Zamini, M., Shukla, D.: Delta-llava: Base-then-specialize alignment for token-efficient vision-language models. In: Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision. pp. 3648–3657 (2026)

62. Zeng, Q.S., Li, Y., Wang, Q., Jiang, P.T., Wu, Z., Cheng, M.M., Hou, Q.: A glimpse to compress: Dynamic visual token pruning for large vision-language models. arXiv preprint arXiv:2508.01548 (2025)
63. Zhang, Q., Liu, M., Li, L., Lu, M., Zhang, Y., Pan, J., She, Q., Zhang, S.: Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. arXiv preprint arXiv:2506.10967 (2025)
64. Zhang, Y., Fan, C.K., Ma, J., Zheng, W., Huang, T., Cheng, K., Gudovskiy, D.A., Okuno, T., Nakata, Y., Keutzer, K., et al.: Sparsevlm: Visual token sparsification for efficient vision-language model inference. In: Forty-second International Conference on Machine Learning (2025)
65. Zhuang, X., Zhu, Z., Xie, Y., Liang, L., Zou, Y.: Vasparse: Towards efficient visual hallucination mitigation via visual-aware token sparsification. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 4189–4199 (June 2025)