PRUNE REDUNDANCY, PRESERVE ESSENCE: VISION TOKEN COMPRESSION IN VLMS VIA SYNERGISTIC IMPORTANCE-DIVERSITY

Anonymous authors
Paper under double-blind review

ABSTRACT

Vision-language models (VLMs) face significant computational inefficiencies caused by excessive generation of visual tokens. While prior work shows that a large fraction of visual tokens are redundant, existing compression methods struggle to balance *importance preservation* and *information diversity*. To address this, we propose PRUNESID, a training-free Synergistic Importance-Diversity approach featuring a two-stage pipeline: (1) Principal Semantic Components Analysis (PSCA) for clustering tokens into semantically coherent groups, ensuring comprehensive concept coverage, and (2) Intra-group Non-Maximum Suppression (NMS) for pruning redundant tokens while preserving key representative tokens within each group. Additionally, PRUNESID incorporates an information-aware dynamic compression ratio mechanism that optimizes token compression rates based on image complexity, enabling more effective average information preservation across diverse scenes. Extensive experiments demonstrate state-of-the-art performance, achieving **96.3%** accuracy on LLaVA-1.5 with only **11.1%** token retention, and **92.8%** accuracy at extreme compression rates (**5.6%**) on LLaVA-NeXT, outperforming prior methods by **2.5%** with **7.8× faster** prefilling speed compared to the original model. Our framework generalizes across diverse VLMs and both image and video modalities, showcasing strong cross-modal versatility.

1 INTRODUCTION

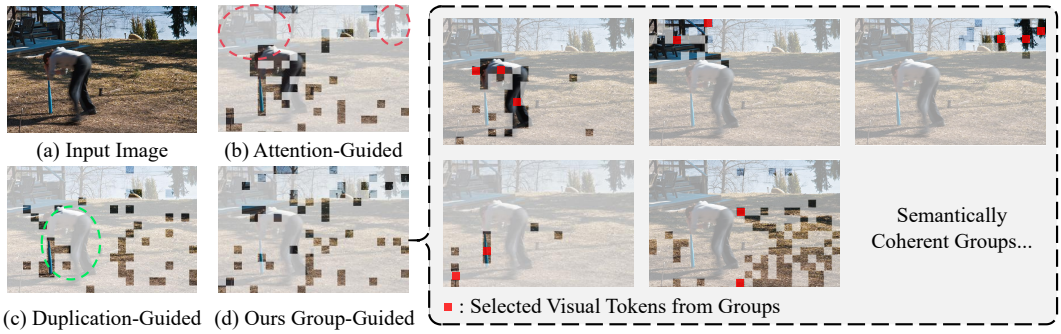

Figure 1: Comparison of visual token reduction paradigms in VLMs. (a) Original input image. (b) Attention-guided methods preserve high-attention tokens but discard contextual background. (c) Duplication-aware methods remove redundant tokens via similarity pruning, yet may discard semantically important regions with high attention. (d) Our proposed semantically group-guided method balances semantic importance and information diversity.

Building upon the success of large language models (LLMs) Brown et al. (2020); Achiam et al. (2023); Touvron et al. (2023), vision-language models (VLMs) Bai et al. (2023); Wang et al. (2024); Wu et al. (2024); Yao et al. (2024); Li et al. (2024) have emerged as a powerful paradigm for multimodal reasoning by encoding images into sequences of visual tokens, thereby enabling joint linguistic and visual understanding. However, this approach introduces substantial computational inefficiencies: contemporary VLMs such as LLaVA-1.5 Liu et al. (2023) and LLaVA-NeXT Liu et al.

(2024b) typically generate 576 and 2880 visual tokens per image, far exceeding what is necessary to capture the essential semantic content of the image. While empirical study Chen et al. (2024a) demonstrates that approximately 70% of visual tokens can be discarded with negligible accuracy degradation, existing compression methodologies fail to optimally reconcile the dual objectives of *importance-aware selection* and *information diversity* at high compression ratios (e.g., retaining only about 5%-10% of tokens), significantly limiting their practical utility for general-purpose VLM applications.

Current visual token reduction techniques can be broadly classified into two paradigms, each exhibiting distinct limitations: Attention-guided selection methods retain visual tokens based on their attention scores Arif et al. (2025); Yang et al. (2024); Zhang et al. (2024b). While effective at preserving semantically salient regions, these approaches systematically neglect contextual background information, thereby compromising scene comprehension. This paradigm suffers from two critical shortcomings: (i) redundant token retention, wherein multiple high-attention patches capture visually similar object segments, inefficiently allocating model capacity to duplicated content; and (ii) contextual degradation, as illustrated in Fig. 1 (b), where the lack of attention to background regions leads to incomplete scene information and weaker overall understanding.

Duplication-aware approaches, exemplified by DART Wen et al. (2025) and DivPrune Alvar et al. (2025), address redundancy through similarity-based pruning. However, these methods exhibit a fundamental limitation: the pruning process inadequately considers token-level semantic importance. Consequently, they may fail to retain tokens with high attention scores that are semantically critical, potentially resulting in incomplete or distorted feature representations, as shown in Fig. 1 (c). These observations reveal an inherent trade-off in token compression: *attention-guided methods preserve local salience at the expense of information diversity, while duplication-aware approaches improve diversity while sacrificing salience preservation.*

To address these limitations, we introduce group-guided PRUNESID, an efficient, generic, and training-free framework that achieves task-agnostic token compression while simultaneously optimizing for both importance preservation and information diversity. Our solution employs a novel two-stage pipeline: (1) Principal Semantic Components Analysis (PSCA), which leverages PCA-driven decomposition Abdi & Williams (2010) to automatically cluster tokens into multiple semantically coherent groups, ensuring comprehensive coverage of critical visual concepts; and (2) Intra-group Non-Maximum Suppression (NMS), which adaptively prunes redundant tokens within each group using dynamic pairwise similarity thresholds (inspired by object detection NMS Neubeck & Van Gool (2006)) while preserving the most semantically significant representatives, as illustrated in Fig. 1 (d). This dual-stage mechanism fundamentally resolves the core trade-off between concept coverage and information density that plagues existing approaches.

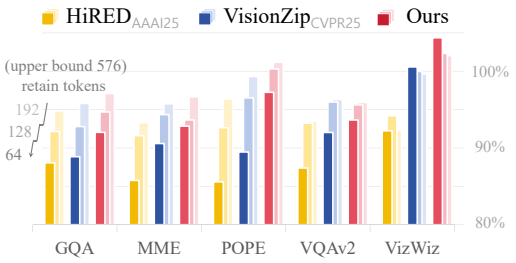

Figure 2: Performance comparison of token reduction methods across multiple vision-language benchmarks on LLaVA-1.5.

Furthermore, PRUNESID incorporates an information-aware dynamic compression ratio mechanism that optimally distributes the token budget per image based on content complexity. This innovation addresses a key limitation of static compression methods by automatically adapting to varying visual semantics, from dense, cluttered scenes to sparse, uniform backgrounds. Our method computes an image-level information score from global token similarity distributions, allocating more tokens to semantically rich images while applying stronger compression to simpler ones. Crucially, this adaptive strategy significantly enhances average information preservation for datasets with high inter-image variability, thereby improving overall model performance.

As demonstrated in Fig. 2, PRUNESID establishes new state-of-the-art performance across multiple vision-language architectures and tasks. The framework achieves 96.3% accuracy on LLaVA-1.5 Liu et al. (2023) while using only 64 tokens (11.1% retention), surpassing VisionZip (92.5%) and HiRED (87.9%) by significant margins. Remarkably, at extreme compression rates (5.6% tokens), it maintains 92.8% accuracy on LLaVA-NeXT Liu et al. (2024b), representing a 2.5 percentage point improvement over prior approaches. Furthermore, PRUNESID demonstrates exceptional scalability by achieving

new SOTA results on Video-LLaVA Lin et al. (2023) with merely 6.6% token retention, confirming its efficacy for both image and video modalities.

In summary, our work makes three principal contributions:

- We propose a training-free visual token compression framework in VLMs that resolves the importance–diversity trade-off via a two-stage pipeline: PSCA for semantic clustering and intra-group NMS for redundancy pruning.
- We introduce an information-aware dynamic compression ratio mechanism that computes a global image-level information score to dynamically assign token budgets across images, enabling effective information preservation in both cluttered and simple scenes.
- Extensive experiments show that our method outperforms prior state-of-the-art across multiple VLMs and tasks, achieving up to 2.5% accuracy gains at extreme compression rates (e.g., 5.6% retention), with strong generalization to image and video modalities.

2 Visual Token Reduction in VLMs

Recent works have identified significant redundancy among visual tokens in VLMs, motivating a line of research focused on training-free methods to improve inference efficiency. A group of approaches Zhang et al. (2024b); Chen et al. (2024b); Wen et al. (2025); Liu et al. (2024c); Dhouib et al. (2025); Yang et al. (2025) conducts token pruning in the early layers of LLMs by leveraging attention-based heuristics. For example, SparseVLM Zhang et al. (2024b) retains visual tokens that receive high average attention scores from textual tokens, indicating stronger textual relevance. Similarly, FastV Chen et al. (2024b) keeps tokens that receive high attention from other tokens, assuming they carry critical information. DART Wen et al. (2025) computes pairwise similarities among tokens and prunes highly similar ones, aiming to retain a less redundant token set. While effective to some extent, these methods still require full token processing in the early LLM layers, incurring non-negligible computational overhead.

To further enhance efficiency, some methods apply token compression in the vision encoder stage, performing *early compression* before interfacing with the LLM. LLaVa-PruMerge Shang et al. (2024) proposes an adaptive selection strategy that leverages the sparsity of attention between the CLS token and visual tokens. It selects tokens with high attention, clusters them based on key similarity, and merges them to enhance information density. HiRED Arif et al. (2025) introduces a hierarchical strategy that partitions the image and allocates a token budget to each region based on CLS attention, enabling a more spatially balanced selection of informative tokens. VisionZip Yang et al. (2024) first identifies dominant tokens with strong attention signals and further merges them based on similarity, ensuring the retention of both salient and contextually rich tokens. These methods significantly reduce the input size to the LLM while maintaining competitive performance.

3 Our Method

3.1 Overview

Given an input image, a pre-trained vision encoder in VLMs first generates a sequence of visual token embeddings, denoted as $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_T\} \in \mathbb{R}^{T \times D}$, where T represents the number of tokens and D denotes the embedding dimension. We aim to reduce this token sequence to a compact representation $\widetilde{\mathbf{X}} \in \mathbb{R}^{N \times D}$ with $N \ll T$, while ensuring: (1) maximal preservation of semantically salient visual patterns and (2) maintaining near-complete information integrity for downstream language modeling tasks.

As illustrated in Fig. 3, we present a novel training-free framework for visual token compression in VLMs. Our methodology employs a two-stage processing pipeline: (1) *semantic-aware token grouping via Principal Semantic Component Analysis (PSCA)*, followed by (2) *intra-group redundancy elimination through Non-Maximum Suppression (NMS)*. The PSCA mechanism clusters tokens by their contribution to semantic principal component directions, generating groups that maintain both semantic coherence and structural diversity. When integrated with adaptive intra-group pruning, this architecture retains compact yet expressive token representation sets that effectively balance information preservation and diversity.

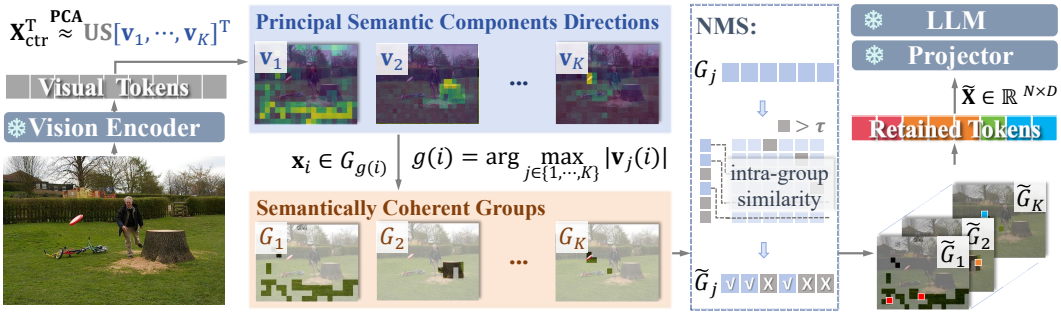

Figure 3: **Overview of our two-stage compression framework.** PSCA first clusters visual tokens into semantically coherent groups via low-rank PCA decomposition. Then, intra-group NMS removes redundant tokens within each group using adaptive similarity thresholds τ, retaining the most informative representatives.

3.2 Semantic-Aware Token Grouping via PSCA

Unlike conventional PCA Abdi & Williams (2010), which operates in the feature dimension to capture variance-driven directions, Principal Semantic Components Analysis (PSCA) redefines the decomposition objective: it models the *token dimension itself* as the semantic axis of interest. By analyzing cross-token variation, PSCA identifies global semantic directions that reflect coherent visual concepts rather than raw statistical variance. This reframing allows PSCA to uncover latent conceptual structures—such as objects, backgrounds, or texture patterns—embedded in the token space.

Specifically, given the token embedding matrix $\mathbf{X}$, we first rescale each element via a sigmoid activation σ to ensure bounded and comparable feature scales. We then center the features across the token dimension to remove global bias. The resulting mean-centered feature matrix is defined as:

$$\mathbf{X}_{\text{ctr}} = \sigma(\mathbf{X}) - \mu, \quad \text{where} \quad \mu = \frac{1}{T}\sum_{i=1}^{T}\sigma(\mathbf{x}_i) \tag{1}$$

so that $\mathbf{X}_{\text{ctr}} \in \mathbb{R}^{T\times D}$ is the zero-mean token matrix, where each row corresponds to one token. We then apply low-rank PCA decomposition to its transpose matrix $\mathbf{X}_{\text{ctr}}^{\top}$:

$$\mathbf{X}_{\text{ctr}}^{\top} \approx \mathbf{U}\mathbf{S}\mathbf{V}^{\top}, \tag{2}$$

where $\mathbf{V} \in \mathbb{R}^{T\times K}$ contains the top-K right singular vectors that define an orthonormal basis over the token dimension. The columns of $\mathbf{V}$ as $\{\mathbf{v}_1, \ldots, \mathbf{v}_K\}$ represent the principal directions of the components. Each row $|\mathbf{V}_{i,:}|$ indicates how much the i-th token contributes to each of the K principal components. A larger value means the token is more strongly related to that component's direction. To form discrete token groups, we assign each token $\mathbf{x}_i$ to the principal direction with the largest absolute value:

$$g(i) = \arg\max_{j} |\mathbf{V}_{i,j}| \,. \tag{3}$$

This procedure partitions the original T tokens into K semantically coherent groups $\{G_1, \ldots, G_K\}$, each capturing shared semantic information across the image.

3.3 Intra-Group Redundancy Removal via NMS

The tokens within each group frequently exhibit significant spatial or semantic overlap, particularly in regions containing dense textures or salient objects. To mitigate this redundancy, we employ a non-maximum suppression (NMS) strategy for each group G_k, which selectively preserves the most informative tokens while eliminating those demonstrating spatial or semantic redundancy.

Following Eq. 3, each token $\mathbf{x}_i \in \mathbf{X}$ is assigned a selection score $s_i = |V_{i,g(i)}|$, which quantifies its contribution to the principal direction of its assigned group $G_{g(i)}$. We then implement greedy NMS within each group as follows: (1) tokens are ranked by their s_i values, and (2) a token is preserved only if its maximum similarity to all previously selected tokens in G_k falls below a threshold τ.

This process generates a refined subset $\widetilde{G}_k \subseteq G_k$ that effectively eliminates redundant tokens while maintaining the original semantic diversity of the group.

To adaptively tune the suppression threshold to varying levels of global redundancy, we introduce a redundancy score ρ defined as the average pairwise similarity among all tokens in the image:

$$\rho = \frac{2}{T(T-1)} \sum_{i=1}^{T} \sum_{j=i+1}^{T} \text{sim}(\mathbf{x}_i, \mathbf{x}_j) \tag{4}$$

where the similarity is computed between ℓ_2-normalized tokens:

$$\text{sim}(\mathbf{x}_i, \mathbf{x}_j) = \frac{\mathbf{x}_i^\top \mathbf{x}_j}{\|\mathbf{x}_i\| \cdot \|\mathbf{x}_j\|}, \quad \forall\, \mathbf{x}_i, \mathbf{x}_j \in \mathbf{X}. \tag{5}$$

We then set the NMS threshold as $\tau = \lambda \cdot \rho$, where λ is a scaling factor determined by the global token budget N. In our experiments, we empirically set $\lambda = \frac{N}{32}$, which consistently worked well across different compression settings. This adaptive threshold encourages stronger suppression for more redundant images.

After performing NMS for all groups, we obtain a collection of filtered groups $\{\widetilde{G}_1, \ldots, \widetilde{G}_K\}$. To match a global token budget N, we allocate group-wise quotas $\{n_1, \ldots, n_K\}$ such that $\sum_{k=1}^{K} n_k = N$, where n_k is calculated by rounding $\frac{|\widetilde{G}_k|}{\sum_j |\widetilde{G}_j|} \cdot N$ to the nearest integer.

Finally, we take the top-n_k tokens from each $\widetilde{G}_k$ according to their selection scores s_i and concatenate them to form the final compact token set:

$$\widetilde{\mathbf{X}} = \bigcup_{k=1}^{K} \text{Top}_{n_k}(\widetilde{G}_k). \tag{6}$$

3.4 Information-Aware Dynamic Compression Ratio Across Images

Conventional token compression methods employ a fixed token compression ratio $r = \frac{N}{T}$ for all images. This uniform approach leads to suboptimal compression: for complex scenes, the predetermined N proves insufficient, causing excessive information loss; whereas for simple scenes, the same N becomes unnecessarily large, resulting in substantial redundancy.

To address this limitation, we propose an *information-aware dynamic compression ratio strategy* that automatically adjusts the retained token budget N according to each image's information content. Building upon the global redundancy measure ρ from Eq. 4, we first compute an image information score:

$$\phi = 1 - \rho, \tag{7}$$

where higher ϕ indicates greater semantic diversity and less redundancy. We then allocate the retained token count N' for each image in proportion to its information score: $N' = \propto \phi$. This ensures that more informative images are allocated more tokens, while simpler images are compressed more aggressively, thereby improving compression adaptiveness across diverse scenes.

4 Experiments

Following the experimental protocol of Yang et al. (2024), we assess the effectiveness of our approach on LLaVA-1.5 Liu et al. (2023). To evaluate generalization, we extend our study to high-resolution vision-language models, including LLaVA-NeXT Liu et al. (2024b) and Mini-Gemini Li et al. (2024). We also conduct experiments on Qwen2-VL Wang et al. (2024) in the supplementary material. Evaluations are conducted using LMMs-Eval Zhang et al. (2024a) on a comprehensive suite of widely used visual understanding benchmarks, including GQA Hudson & Manning (2019), MMBench Liu et al. (2024d), MME Fu et al. (2023), POPE Li et al. (2023b), ScienceQA Lu et al. (2022), VQA-v2 Goyal et al. (2017), TextVQA Singh et al. (2019), MMMU Yue et al. (2024), SEED-Bench Li et al. (2023a), VizWiz Gurari et al. (2018), and LLaVA-Bench Liu et al. (2024a). We further evaluate the applicability of our method to video understanding tasks using Video-LLaVA Lin et al. (2023).

Table 1: **Performance of PRUNESID on LLaVA-1.5.** *Vanilla* refers to the uncompressed baseline model using all 576 visual tokens, serving as the upper performance bound. *Early Cmp.* indicates whether the compression is applied prior to the LLM for improved efficiency. PRUNESID-Dyn denotes the variant of out method augmented with the Dynamic Compression Ratio mechanism.

Method	Early Cmp.	GQA	MMB	MME	POPE	SQA	VQA^{v2}	VQA^{Text}	MMMU	SEED	VizWiz	LLaVa-B	Avg.
Upper Bound, 576 Tokens **(100%)**													
Vanilla CVPR24	–	61.9	64.7	1862	85.9	69.5	78.5	58.2	36.3	60.5	54.3	66.8	100%
Retain 192 Tokens (↓ **66.7%**)													
FastV ECCV24	×	52.7	61.2	1612	64.8	67.3	67.1	52.5	34.3	57.1	50.8	49.4	88.2%
SparseVLM 24.10	×	57.6	62.5	1721	83.6	69.1	75.6	56.1	33.8	55.8	50.5	66.1	95.3%
MustDrop 24.11	×	58.2	62.3	1787	82.6	69.2	76.0	56.5	–	–	51.4	–	96.3%
DART 25.02	×	60.0	63.6	1856	82.8	69.8	76.7	57.4	36.4	51.5	54.9	64.2	97.3%
ToMe ICLR23	✓	54.3	60.5	1563	72.4	65.2	68.0	52.1	–	–	–	–	88.5%
LLaVa-PruMerge 24.05	✓	54.3	59.6	1632	71.3	67.9	70.6	54.3	–	–	50.1	–	90.5%
HiRED AAAI25	✓	58.7	62.8	1737	82.8	68.4	74.9	47.4	–	–	50.1	–	93.6%
DivPrune CVPR25	✓	60.0	62.3	1752	87.0	68.7	75.5	56.4	35.8	58.6	55.6	64.8	96.9%
VisionZip CVPR25	✓	59.3	63.0	1783	85.3	68.9	76.8	57.3	36.3	58.5	54.1	67.7	98.4%
VisionZip -Dyn	✓	59.4	63.3	1797	85.5	68.7	76.9	57.4	36.8	58.6	54.4	67.7	98.6%
PRUNESID	✓	60.1	63.7	1791	86.9	68.5	76.8	56.7	36.1	59.0	55.4	65.1	**98.5%**
PRUNESID-Dyn	✓	60.2	63.8	1797	87.1	69.1	76.8	56.9	36.8	59.0	55.5	65.1	**98.6%**
Retain 128 Tokens (↓ **77.8%**)													
FastV ECCV24	×	49.6	56.1	1490	59.6	60.2	61.8	50.6	34.9	55.9	51.3	52.0	84.5%
SparseVLM 24.10	×	56.0	60.0	1696	80.5	67.1	73.8	54.9	33.8	53.4	51.4	62.7	93.0%
MustDrop 24.11	×	56.9	61.1	1745	78.7	68.5	74.6	56.3	–	–	52.1	–	94.7%
DART 25.02	×	58.7	63.2	1840	80.1	69.1	75.9	56.4	36.2	50.5	55.3	62.4	96.0%
ToMe ICLR23	✓	52.4	53.3	1343	62.8	59.6	63.0	49.1	–	–	–	–	80.4%
LLaVa-PruMerge 24.05	✓	53.3	58.1	1554	67.2	67.1	68.8	54.3	–	–	50.3	–	88.9%
HiRED AAAI25	✓	57.2	61.5	1710	79.8	68.1	73.4	46.1	–	–	51.3	–	92.2%
DivPrune CVPR25	✓	59.2	62.3	1752	86.9	69.0	74.7	56.0	36.2	57.1	55.6	66.2	96.1%
VisionZip CVPR25	✓	57.6	62.0	1762	83.2	68.9	75.6	56.8	37.9	57.1	54.5	64.8	97.2%
VisionZip -Dyn	✓	57.6	62.2	1770	83.5	68.9	75.8	56.9	37.5	57.8	54.7	65.3	97.5%
PRUNESID	✓	58.8	62.1	1749	86.5	68.3	75.3	54.7	35.8	57.8	55.8	68.8	**97.6%**
PRUNESID-Dyn	✓	58.9	62.6	1760	86.9	68.8	75.4	55.1	36.3	57.9	56.0	68.9	**98.1%**
Retain 64 Tokens (↓ **88.9%**)													
FastV ECCV24	×	46.1	48.0	1256	48.0	51.1	55.0	47.8	34.0	51.9	50.8	46.1	76.3%
SparseVLM 24.10	×	52.7	56.2	1505	75.1	62.2	68.2	51.8	32.7	51.1	53.1	57.5	87.6%
MustDrop 24.11	×	53.1	60.0	1612	68.0	63.4	69.3	54.2	–	–	51.2	–	88.9%
DART 25.02	×	55.9	60.6	1765	73.9	69.8	72.4	54.4	35.9	47.2	55.3	59.1	92.6%
ToMe ICLR23	✓	48.6	43.7	1138	52.5	50.0	57.1	45.3	–	–	–	–	70.1%
LLaVa-PruMerge 24.05	✓	51.9	55.3	1549	65.3	68.1	67.4	54.0	–	–	50.1	–	87.2%
HiRED AAAI25	✓	54.6	60.2	1599	73.6	68.2	68.7	44.2	–	–	50.2	–	88.4%
DivPrune CVPR25	✓	57.6	59.3	1638	85.6	68.3	72.9	55.5	36.3	55.4	57.5	64.0	94.6%
VisionZip CVPR25	✓	55.1	60.1	1690	77.0	69.0	72.4	55.5	36.2	54.5	54.8	62.9	94.0%
VisionZip -Dyn	✓	55.2	60.1	1694	77.1	69.2	72.8	55.8	36.7	54.7	54.9	63.1	94.4%
PRUNESID	✓	57.1	58.8	1733	83.8	67.8	73.7	54.2	37.0	56.1	56.9	65.2	**95.9%**
PRUNESID-Dyn	✓	57.2	59.7	1734	84.1	68.1	73.8	54.2	37.2	56.2	57.0	65.8	**96.3%**

4.1 MAIN RESULTS ON IMAGE UNDERSTANDING TASKS

Results on LLaVA-1.5. LLaVA-1.5 uniformly resizes input images to a resolution of 336×336 before passing them through a CLIP-based Radford et al. (2021) vision encoder, which produces 576 visual tokens. Following prior work Chen et al. (2024b); Zhang et al. (2024b); Yang et al. (2024), we conduct experiments under three token retention settings: 64, 128, and 192 tokens. As shown in Tab. 1, our method consistently achieves state-of-the-art average performance across all configurations, outperforming both early-stage compression approaches that apply compression during the image encoder stage and more computationally intensive methods applied during the prefilling stage. Notably, when retaining only 64 image tokens—equivalent to merely 11.1% of the original token count—our method achieves an average accuracy of approximately **96%** across all benchmarks, surpassing the strong prior method VisionZip by a margin of **1.9%**. This result highlights the superior information richness of the visual tokens selected by our method under extreme compression settings.

PRUNESID-Dyn and VisionZip-Dyn denote the variant of our method and VisionZip augmented with the information-aware dynamic compression ratio mechanism (Sec. 3.4). To ensure a fair comparison, we constrain the average number of retained tokens per benchmark to match that of the fixed-budget setting. Experimental results show that the dynamic strategy consistently improves performance. Notably, its effectiveness varies across benchmarks, which we further analyze in detail in Sec. 4.3.

Results on LLaVA-NeXT. LLaVA-NeXT Liu et al. (2024b) divides the image into multiple parts based on its aspect ratio for vision encoding, resulting in a maximum sequence length of up to 2880 tokens. (i.e., 576 tokens × 5). Following the evaluation protocol in Yang et al. (2024), we assess

Table 2: Performance on LLaVA-NeXT.

Method	GQA	MMB	MME	POPE	SQA	VQA^{v2}	MMMU	SEED^{I}	Avg.
Upper Bound, 2880 Tokens (**100%**)									
Vanilla	64.2	67.9	1842	86.4	70.2	80.1	35.1	70.2	100%
Retain 640 Tokens (↓ **77.8%**)									
VisionZip	61.3	66.3	1787	86.3	68.1	79.1	34.7	66.7	97.5%
PRUNESID	61.6	64.2	1795	86.3	68.3	78.5	37.9	67.3	**98.4%**
Retain 320 Tokens (↓ **88.9%**)									
VisionZip	59.3	63.1	1702	82.1	67.3	76.2	35.3	63.4	94.3%
PRUNESID	60.5	63.0	1754	83.1	67.3	76.6	36.4	65.0	**95.8%**
Retain 160 Tokens (↓ **94.4%**)									
VisionZip	55.5	60.1	1630	74.8	68.3	71.4	36.1	58.3	90.3%
PRUNESID	58.9	60.8	1704	76.9	67.1	73.8	36.2	62.5	**92.8%**

Table 3: Performance on Mini-Gemini.

Method	GQA	MMB	MME	POPE	SQA	VQA^{v2}	MMMU	SEED^{I}	Avg.
Upper Bound, 576 Tokens (**100%**)									
Vanilla	62.4	69.3	1841	85.8	70.7	80.4	36.1	69.7	100%
Retain 192 Tokens (↓ **66.7%**)									
VisionZip	60.3	68.9	1846	82.3	70.1	79.1	36.1	67.5	98.3%
PRUNESID	61.2	67.2	1842	84.4	71.1	79.1	36.1	67.8	**98.7%**
Retain 128 Tokens (↓ **77.8%**)									
VisionZip	58.7	68.1	1841	78.5	70.0	77.5	34.8	65.6	96.2%
PRUNESID	60.1	66.6	1821	82.4	70.7	77.8	36.0	66.5	**97.4%**
Retain 64 Tokens (↓ **88.9%**)									
VisionZip	55.8	65.9	1737	69.6	70.7	73.9	35.6	61.7	92.4%
PRUNESID	58.3	63.1	1735	76.0	70.6	75.2	37.2	63.6	**94.4%**

our method under three token retention ratios: 22.2%, 11.1%, and 5.6% of the total visual tokens. The results are presented in Tab. 2. Compared to the strong prior method VisionZip, our approach achieves average performance gains of 0.9%, 1.5%, and **2.5**% under the above three token retention settings, respectively. Notably, even when retaining only about 5% of the original image tokens, our method enables the vision-language model to preserve **92.8%** of its full performance, demonstrating its ability to maximize information preservation without introducing task-specific biases—such as overemphasizing foreground content at the expense of contextual or background information.

Results on Mini-Gemini. Following Yang et al. (2024), we also evaluate the generalizability of our method on the Mini-Gemini model to demonstrate its effectiveness across diverse VLM architectures. Mini-Gemini incorporates a high-resolution vision encoder based on ConvNeXt-L Liu et al. (2022) to extract fine-grained visual features. We apply our token compression method to the final image tokens produced by the vision encoder and evaluate the model's inference performance under various token retention settings across multiple benchmarks. As shown in Tab. 3, our method consistently delivers strong performance, validating its robustness across architectures with different vision backbones.

4.2 Main results on Video Understanding Tasks

To further evaluate the effectiveness of our method on video understanding tasks, we apply PRUNESID to Video-LLaVA Lin et al. (2023) and conduct experiments on four video question answering benchmarks: TGIF Jang et al. (2017), MSVD Xu et al. (2017), MSRVTT Xu et al. (2017), and ActivityNet Yu et al. (2019). Each input video consists of 8 frames, with 256 tokens per frame, resulting in 2048 tokens.

Table 4: Performance on Video-LLaVA.

Method	TGIF	MSVD	MSRVTT	A-Net	Avg.
Video-LLaVA	47.1	70.7	59.2	43.1	100%
FastV	23.1	38.0	19.3	30.6	52.1%
SparseVLM	44.7	68.2	31.0	42.6	86.5%
VizionZip	42.4	63.5	52.1	43.0	93.2%
PRUNESID	45.8	67.1	53.3	43.1	**95.5%**

Following prior works Chen et al. (2024b); Zhang et al. (2024b); Yang et al. (2024), we compress 256 tokens of each frame into 17 tokens, retaining only 6.6%. This reduces the full 2048 video tokens to just 136, which are then passed to the subsequent stages. As shown in Tab. 4, our method achieves consistently better performance than strong prior approaches across all benchmarks, reaching an average accuracy of **95.5%**. These results highlight the strength of our synergistic importance-diversity approach in preserving key semantically representative information under high compression ratios, thereby enabling stronger generalization across diverse video understanding tasks.

4.3 Ablation Study

Ablation on Token Grouping Strategy. We evaluate the effect of different token grouping strategies used before the intra-group NMS. Specifically, we compute our PSCA-based grouping method with two alternatives: i) a *random grouping* baseline where tokens are shuffled and uniformly partitioned into groups, and ii) a *KMeans-based grouping* Lloyd (1982) applied directly on the token features. As shown in Tab. 5, PSCA consistently outperforms other methods across four benchmarks. This demonstrates the advantage of PSCA in forming semantically coherent token groups by leveraging the local principal subspace structure, leading to more effective redundancy reduction.

Table 5: Ablation study of group method on LLaVA-1.5.

Method	Retain 64				Retain 128				Retain 192				Avg.
	GQA	MME	POPE	SQA	GQA	MME	POPE	SQA	GQA	MME	POPE	SQA	
random	56.2	1707	79.6	67.5	57.8	1723	84.0	66.0	59.4	1743	85.6	67.5	94.8%
kmeans	56.5	1630	82.8	67.8	58.7	1714	86.3	67.8	60.0	1745	86.8	68.0	95.6%
PRUNESID	57.1	1733	83.8	67.8	58.8	1749	86.5	68.3	60.1	1793	86.9	68.5	**96.8%**

Table 6: Ablation study of group K on LLaVA-1.5.

K	Retain 64					Retain 128					Retain 192				
	GQA	MME	POPE	SQA	Avg.	GQA	MME	POPE	SQA	Avg.	GQA	MME	POPE	SQA	Avg.
8	56.2	1684	81.4	67.1	93.1%	58.2	1720	85.3	68.3	96.0%	59.2	1763	83.0	67.9	96.2%
16	57.1	1733	83.8	67.8	**95.1%**	58.6	1692	86.4	68.2	96.1%	59.2	1755	85.4	68.3	96.9%
32	57.1	1700	83.7	68.0	94.7%	58.8	1749	86.5	68.3	**97.0%**	60.1	1763	86.8	68.4	97.9%
48	56.9	1668	83.7	67.6	94.1%	58.7	1720	85.8	68.3	96.3%	60.1	1793	86.9	68.5	**98.3%**
60	56.7	1657	83.6	67.6	93.8%	58.6	1715	85.7	68.0	96.1%	60.0	1774	86.9	68.5	98.0%

Ablation on Token Group Counts K. We study how the number of token groups (K) affects PRUNESID performance under different total retained token counts ($N \in \{64, 128, 192\}$). For each N, we sweep the number of groups K from 8 to 64. As shown in Tab. 6, performance exhibits a bell-shaped trend: too few groups reduce the granularity of redundancy modeling, while too many groups lead to overly small group sizes and unstable pruning. The optimal settings align with moderate values of $K = \frac{N}{4}$. These results validate our heuristic choice of increasing K proportionally with N, balancing diversity and intra-group competition.

Ablation on ViT Layer Features for PSCA Grouping. We analyze the effect of using features from different ViT layers for PSCA grouping. As Fig. 4 shows, middle-to-late layers (16, 22) yield better results across multiple metrics, indicating more effective semantic clustering. Early layers (0, 2) underperform due to weaker semantic information. Notably, the final output layer (23) shows a slight drop or plateau compared to layer 22, likely because layer 22 features are directly used for LLM training and thus better capture the semantic information needed for token grouping, whereas layer 23's final output is more specialized and less balanced for this purpose. These results validate our choice to extract intermediate-late layer features (e.g., layer 22) for PSCA, striking a balance between semantic richness and balanced coverage.

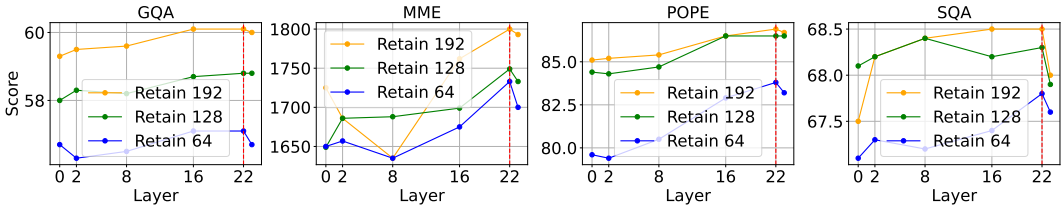

Figure 4: Ablation study on ViT layer features for PSCA Grouping.

Ablation on Dynamic Compression Ratio Mechanism. We have demonstrated the effectiveness of the dynamic compression ratio mechanism in Tab. 1. Furthermore, we conduct an in-depth analysis of its performance variations across diverse benchmarks and its generalization capability across multiple model architectures. Theoretically, as the heterogeneity of information scores among test images increases, the adaptive adjustment capacity of the dynamic compression ratio mechanism broadens, thereby amplifying performance enhancements. This hypothesis is corroborated by the distribution depicted in Fig. 5, where the MMMU benchmark demonstrates significantly greater information score variability relative to GQA—a trend consistent with the enhanced performance gains observed for MMMU in Tab. 1.

To further validate the advantages of the dynamic compression ratio strategy, we conduct comprehensive experiments on LLaVA-1.5, LLaVA-NeXT, and Mini-Gemini across diverse datasets characterized by high information variance, including MME, ScienceQA, MMMU, and POPE. As shown in Tab. 7, our dynamic strategy consistently surpasses fixed-token baselines, achieving up to 1.0 % performance improvements under identical average token budgets. These findings underscore the efficacy of adaptive token compression in average information preserving—particularly beneficial for benchmarks with substantial inter-image variability.

Table 7: Ablation study of dynamic compression ratio mechanism.

Methods	LLaVA-1.5						LLaVA-NeXT						Mini-Gemini					
	MME	SQA	MMMU	POPE	Avg.	Δ	MME	SQA	MMMU	POPE	Avg.	Δ	MME	SQA	MMMU	POPE	Avg.	Δ
baseline	1862	69.5	36.3	85.9	100%		1842	70.2	35.1	86.4	100%		1841	70.7	36.1	85.8	100%	
	Retain 192 Tokens (↓ 66.7%)						*Retain 128 Tokens (↓ 77.8%)*						*Retain 192 Tokens (↓ 66.7%)*					
PRUNESID	1791	68.5	36.1	86.9	98.8%		1795	68.3	37.9	86.3	100.7%		1842	71.1	36.1	84.4	99.7%	
PRUNESID-Dyn	1797	69.1	36.8	87.1	99.7%	+0.9	1798	68.9	38.2	86.5	101.2%	+0.5	1845	71.3	36.9	84.6	100.5%	+0.8
	Retain 128 Tokens (↓ 77.8%)						*Retain 64 Tokens (↓ 88.9%)*						*Retain 128 Tokens (↓ 77.8%)*					
PRUNESID	1749	68.3	35.8	86.5	97.9%		1754	67.3	36.4	83.1	97.7%		1821	70.7	36.0	82.4	98.7%	
PRUNESID-Dyn	1760	68.8	36.3	86.9	98.7%	+0.8	1787	67.7	36.8	83.5	98.7%	+1.0	1846	71.3	36.4	82.5	99.5%	+0.8
	Retain 64 Tokens (↓ 88.9%)						*Retain 32 Tokens (↓ 94.4%)*						*Retain 64 Tokens (↓ 88.9%)*					
PRUNESID	1733	67.8	37.0	83.8	97.5%		1704	67.1	36.2	76.9	95.1%		1735	70.6	37.2	76.0	96.4%	
PRUNESID-Dyn	1734	68.1	37.2	84.1	97.9%	+0.4	1744	67.7	36.4	77.4	96.1%	+1.0	1760	70.8	37.2	76.9	97.1%	+0.7

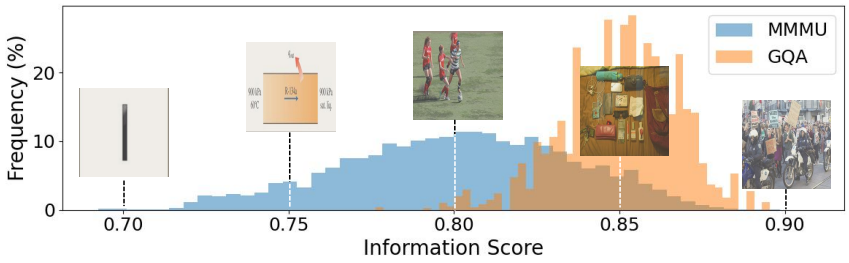

Figure 5: Histogram of Information Score distributions for the MMMU and GQA benchmarks. A higher Information Score indicates greater visual information content.

4.4 Efficiency Analysis.

Vision-language models (VLMs) suffer from prolonged prefilling time due to excessive visual tokens generated by dense image encoding. As shown in Tab. 8, on the POPE benchmark, LLaVA-NeXT 7B produces up to 2,800 visual tokens per image, where prefilling occupies 86% of the total inference time (254 ms/sample).

For consistency with prior work Yang et al. (2024), we report inference time on a single NVIDIA A800-80GB. At a compression rate of 5.6% (retaining only 160 tokens), our approach reduces prefilling time from 218ms to just 27.8ms—a 7.8× improvement—while also decreasing overall inference time to 89ms per sample. Compared to VisionZip, which achieves similar latency (27.8ms prefilling, 84ms inference), our method maintains the same level of efficiency but delivers superior performance on POPE, improving F_1 score from 86.6% to 89.0% (+2.4%). This demonstrates that our method not only preserves computational efficiency but also retains more semantically relevant visual information during compression.

Table 8: **Efficiency analysis and comparison.** Inference and prefilling times represent the average per-sample latency.

Method	Token	Inference Time ↓	Prefilling Time ↓	POPE (F_1)↑
LLaVA-NeXT	2880	254ms	218ms	86.4
FastV	160	199ms	119ms	50.5%
SparseVLM	160	211ms	128ms	80.2%
VisionZip	160	**84ms**	**27.8ms**	86.6%
PRUNESID	160	89ms	**27.8ms**	**89.0%**

5 Conclusion

In this work, we present PRUNESID, a training-free and task-agnostic framework for efficient visual token reduction in vision-language models (VLMs). By integrating Principal Semantic Component Analysis (PSCA) for semantically coherent grouping with intra-group Non-Maximum Suppression (NMS) for redundancy pruning, PRUNESID effectively balances *importance-aware selection* and *information diversity*. Moreover, its dynamic compression ratio mechanism adapts retained token counts based on image complexity, leading to improved overall performance. Extensive experiments demonstrate state-of-the-art results across both image and video VLM benchmarks, retaining ∼5% of visual tokens while achieving 92.8% and 95.5% accuracy on LLaVA-NeXT and Video-LLaVA, respectively. These results highlight the potential of semantically group-guided token selection for scaling VLMs to more demanding and resource-constrained settings.

REFERENCES

Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.

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

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

Kazi Hasan Ibn Arif, JinYi Yoon, Dimitrios S Nikolopoulos, Hans Vandierendonck, Deepu John, and Bo Ji. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1773–1781, 2025.

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

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Jieneng Chen, Luoxin Ye, Ju He, Zhao-Yang Wang, Daniel Khashabi, and Alan Yuille. Efficient large multi-modal models via visual context compression. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024a.

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

Mohamed Dhouib, Davide Buscaldi, Sonia Vanier, and Aymen Shabou. Pact: Pruning and clustering-based token reduction for faster visual language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 14582–14592, 2025.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 6904–6913, 2017.

Danna Gurari, Qing Li, Abigale J Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P Bigham. 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.

Drew A Hudson and Christopher D Manning. 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.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 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.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. Seed-bench: Benchmarking multimodal llms with generative comprehension. *arXiv preprint arXiv:2307.16125*, 2023a.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. Mini-gemini: Mining the potential of multi-modality vision language models. *arXiv preprint arXiv:2403.18814*, 2024.

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

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 26296–26306, 2024a.

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

Ting Liu, Liangtao Shi, Richang Hong, Yue Hu, Quanjun Yin, and Linfeng Zhang. Multi-stage vision token dropping: Towards efficient multimodal large language model. *arXiv preprint arXiv:2411.10803*, 2024c.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024d.

Zhuang Liu, Hanzi Mao, Chao-Yuan Wu, Christoph Feichtenhofer, Trevor Darrell, and Saining Xie. A convnet for the 2020s. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 11976–11986, 2022.

Stuart Lloyd. Least squares quantization in pcm. *IEEE transactions on information theory*, 28(2): 129–137, 1982.

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

Alexander Neubeck and Luc Van Gool. Efficient non-maximum suppression. In *18th international conference on pattern recognition (ICPR'06)*, volume 3, pp. 850–855. IEEE, 2006.

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

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

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

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

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

Zhiyu Wu, Xiaokang Chen, Zizheng Pan, Xingchao Liu, Wen Liu, Damai Dai, Huazuo Gao, Yiyang Ma, Chengyue Wu, Bingxuan Wang, et al. Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. *arXiv preprint arXiv:2412.10302*, 2024.

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 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.

Longrong Yang, Dong Shen, Chaoxiang Cai, Kaibing Chen, Fan Yang, Tingting Gao, Di Zhang, and Xi Li. Libra-merging: Importance-redundancy and pruning-merging trade-off for acceleration plug-in in large vision-language model. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9402–9412, 2025.

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

Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pp. 9127–9134, 2019.

Xiang Yue, Yuansheng Ni, Kai Zhang, Tianyu Zheng, Ruoqi Liu, Ge Zhang, Samuel Stevens, Dongfu Jiang, Weiming Ren, Yuxuan Sun, et al. Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 9556–9567, 2024.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, et al. Lmms-eval: Reality check on the evaluation of large multimodal models. *arXiv preprint arXiv:2407.12772*, 2024a.

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