IDPruner: Harmonizing Importance and Diversity in Visual Token Pruning for MLLMs

Anonymous ACL submission

Abstract

Multimodal Large Language Models (MLLMs) have demonstrated impressive capabilities, yet they encounter significant computational bottlenecks due to the massive volume of visual tokens. Consequently, visual token pruning, which substantially reduces the token count, has emerged as a critical technique for accelerating MLLM inference. Existing approaches focus on token importance, diversity, or an intuitive combination of both, without a principled framework for their optimal integration. To address this issue, we first conduct a systematic analysis to characterize the trade-off between token importance and semantic diversity. Guided by this analysis, we propose the **I**mportance and **D**iversity Pruner (**IDPruner**), which leverages the Maximal Marginal Relevance (MMR) algorithm to achieve a Pareto-optimal balance between these two objectives. Crucially, our method operates without requiring attention maps, ensuring full compatibility with FlashAttention and efficient deployment via one-shot pruning. We conduct extensive experiments across various model architectures and multimodal benchmarks, demonstrating that IDPruner achieves state-of-the-art performance and superior generalization across diverse architectures and tasks. Notably, on Qwen2.5-VL-7B-Instruct, IDPruner retains 95.18% of baseline performance when pruning 75% of the tokens, and still maintains 86.40% even under an extreme 90% pruning ratio. Our code is available at `https://anonymous.4open.science/r/IDPruner-B4B6/`.

1 Introduction

Multimodal Large Language Models (MLLMs) have achieved significant success in artificial intelligence. These models typically encode images or videos into sequences of visual tokens, which are then processed together with textual inputs by the language model to generate text responses (Liu et al., 2023b,a). For instance, Qwen2.5-VL generates approximately 2,691 visual tokens when processing a single 1080p image (1920×1080), with each token representing a 28×28 pixel patch. The high number of visual tokens creates a heavy computational burden, limiting the efficiency and practical deployment of MLLMs (Zhou et al., 2024). Thus, visual token pruning (Wang et al., 2025; Shao et al., 2025), which aims to reduce the number of visual tokens while maintaining model performance, has emerged as a critical technique for achieving efficient MLLM inference.

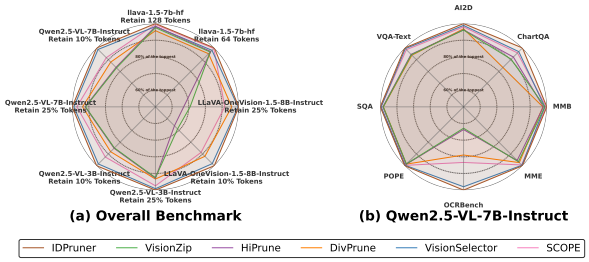

Figure 1: **Performance comparison across four architectures and eight benchmarks.** IDPruner (outermost boundary) consistently outperforms baselines in both **(a) aggregated performance** across four diverse MLLM architectures and **(b) fine-grained benchmark breakdown** for Qwen2.5-VL. This demonstrates the superior cross-architecture generalization and task-specific robustness of our method.

Existing pruning strategies generally fall into two categories: importance-based and diversity-based methods. Importance-based approaches (Chen et al., 2024a; Yang et al., 2025b,a) select salient tokens, focusing on foreground objects, but often sacrificing the background context essential for global reasoning. In contrast, diversity-based methods (Alvar et al., 2025; Zou et al., 2025) maximize semantic coverage to reduce redundancy but risk retaining task-irrelevant noise while missing fine-grained details. Recent hybrid approaches (Zhang et al., 2024c, 2025a; Li et al., 2025) attempt

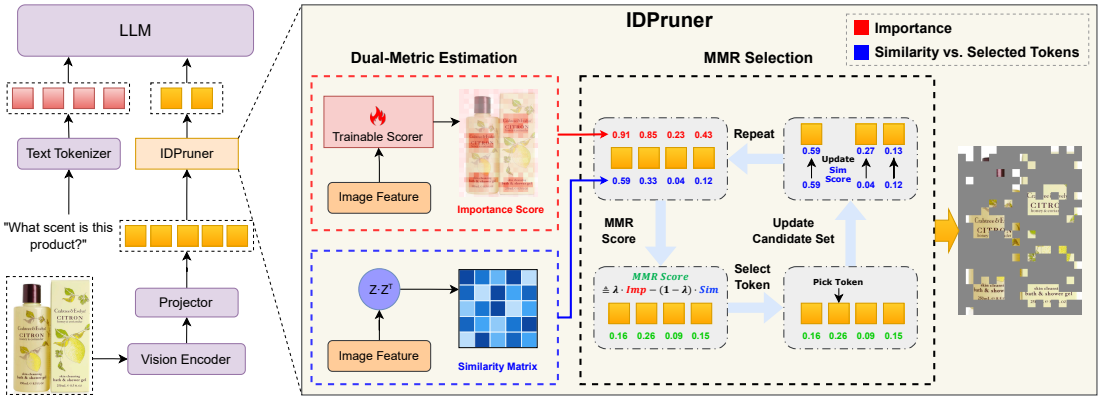

Figure 2: **Overview of the IDPruner framework. Left**: Integration of our one-shot visual token pruning into the MLLM inference pipeline. **Right**: The core mechanism computes **Importance Scores** (Red) and a **Similarity Matrix** (Blue), utilizing an MMR selection process to harmonize importance and diversity. This approach operates without attention maps and remains compatible with FlashAttention.

to combine these complementary criteria but lack rigorous analysis, relying on intuition-based integration that yields suboptimal performance. Therefore, a systematic analytical framework is needed to characterize the interaction between importance and diversity and derive optimal integration strategies.

To address this, we first conduct a systematic analysis to investigate the trade-off between token importance and semantic diversity. As shown in Figure 3, our analysis reveals that current approaches fail to effectively balance these two critical dimensions. To overcome this limitation, we introduce the **I**mportance and **D**iversity Pruner (**IDPruner**), a novel pruning strategy designed to balance these criteria optimally. Specifically, as illustrated in Figure 2, we cast visual token pruning as a re-ranking problem in information retrieval and adapt the Maximal Marginal Relevance (MMR) (Carbonell and Goldstein-Stewart, 1998) algorithm to model the interplay between token importance and semantic diversity explicitly. This approach selects tokens that jointly maximize both importance and diversity.

IDPruner achieves state-of-the-art performance, as demonstrated by comprehensive evaluations on multimodal benchmarks. Notably, on the Qwen2.5-VL-7B-Instruct model, even under an extreme compression ratio of 90%, our method retains **86.40%** of the baseline performance, significantly outperforming existing competitive approaches. Crucially, unlike progressive pruning strategies that dynamically change sequence lengths, IDPruner performs one-shot pruning at an early stage, which makes it easier to integrate into inference engines like vLLM (Kwon et al., 2023). Furthermore, our method works without requiring attention information, ensuring full compatibility with FlashAttention (Dao et al., 2022) to maximize inference efficiency.

The main contributions of this work are summarized as follows:

- We conduct a systematic analysis to characterize the trade-off between token importance and semantic diversity, providing a theoretical basis for their integration.
- We propose IDPruner, which adapts the Maximal Marginal Relevance (MMR) algorithm to visual token pruning, enabling the optimal harmonization of importance and diversity.
- Extensive experiments demonstrate that our method achieves state-of-the-art performance and exceptional cross-architecture generalization, as visualized in Figure 1, while supporting one-shot pruning and FlashAttention acceleration, offering a practical solution for efficient MLLM deployment.

2 Related work

Large Multimodal Models and Visual Token Pruning. Recent Multimodal Large Language Models (MLLMs) (Liu et al., 2023a; Wang et al., 2024; Zhu et al., 2025b) have demonstrated impressive capabilities across various visual tasks, yet they encounter significant computational bottlenecks due to the massive volume of visual tokens. Static-resolution models like LLaVA-1.5 (Liu et al.,

2023a) and LLaVA-NeXT (Liu et al., 2024a) require 576 and 2,880 input tokens per image, respectively, while newer architectures such as the Qwen-VL (Bai et al., 2025), LLaVA-OneVision (Li et al., 2024a), and InternVL (Zhu et al., 2025b) series demand comparable token budgets for high-resolution processing. Consequently, visual token pruning, which eliminates unnecessary tokens, has emerged as a crucial technique for accelerating MLLM inference. Current research typically falls into two categories: importance-based methods and diversity-based methods.

Importance-based Token Pruning. Importance-based approaches reduce computational overhead by retaining only the most salient tokens. Early studies rely on attention scores from LLM decoder layers (Chen et al., 2024a; Zhang et al., 2024e; Xing et al., 2024; Zhang et al., 2025b; Ye et al., 2025; Han et al., 2025), while subsequent research discovers that the attention of the [CLS] token in Vision Transformers (ViT) provides a more effective importance measure (Yang et al., 2025b; Liu et al., 2025; Zhang et al., 2024d; Tong et al., 2025). To mitigate limitations such as FlashAttention incompatibility, recent work has introduced alternative metrics, including optimal transport and L2 norms (Yang et al., 2025a; Dhouib et al., 2025). Beyond training-free methods, approaches like VisionSelector (Zhu et al., 2025a) employ learnable modules to estimate token importance, achieving state-of-the-art performance through end-to-end training. Despite their effectiveness in capturing region-specific details, these methods often overlook global context, potentially causing information loss in background areas.

Diversity-based Token Pruning. Diversity-based approaches aim to preserve information coverage by regarding visual tokens as a collective set, minimizing redundancy to retain a representative subset of visual features. DivPrune (Alvar et al., 2025) formulates this task as a Max-Min Diversity Problem, solving it via a greedy algorithm to maximize semantic coverage, while DART (Wen et al., 2025) employs a parallelizable strategy that selects pivot tokens and eliminates their nearest neighbors to maintain diversity. However, maximizing redundancy reduction often comes at the cost of missing fine-grained details in focal regions, as these methods may indiscriminately retain task-irrelevant noise.

Hybrid Strategies. Synergizing importance and diversity typically yields superior performance compared to single-criterion methods. VisPrune (Zhang et al., 2024c) allocates token budgets based on both [CLS] attention and diversity, while CDPruner (Zhang et al., 2025a) employs Determinantal Point Processes (DPP) to balance these objectives. Other approaches explore alternative integration strategies, such as ensuring spatial coverage via region-based allocation (Zou et al., 2025; Arif et al., 2025) or modeling pruning as a set cover problem to optimize multimodal coverage (Li et al., 2025; Deng et al., 2025). Although effective, these methods typically rely on heuristic integration strategies without a systematic analytical framework. In this work, we address this limitation by introducing a systematic framework that optimally harmonizes importance and diversity.

3 An Empirical Analysis of the Importance-Diversity Trade-off

3.1 Quantifying Importance and Diversity

Visual token pruning strategies typically focus on either importance-based selection or diversity preservation; however, balancing these two goals remains challenging. To systematically analyze the relationship between these two paths, we first reformulate the visual token pruning problem.

Definition 1 (Visual Token Pruning)**.** *Let $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ denote the set of N visual tokens, where each token $v_i \in \mathbb{R}^d$ represents a d-dimensional feature vector. Visual token pruning aims to select a subset $\mathcal{S} \subset \mathcal{V}$ with $|\mathcal{S}| = K < N$ tokens, where K is a pre-defined budget constraint.*

To decouple the combining strategy from any specific importance estimator, we pre-define an importance vector $\mathbf{w}$ representing the weight of each token, regardless of how $\mathbf{w}$ is calculated. Based on this, we define the retention metric:

Definition 2 (Importance Retention Ratio)**.** *The importance retention ratio of a subset $\mathcal{S}$ is defined as the normalized sum of retained scores:*

$$\mathcal{I}(\mathcal{S}) = \frac{\sum_{v_k \in \mathcal{S}} w_k}{\sum_{v_i \in \mathcal{V}} w_i}$$

This metric quantifies the proportion of total information retained by the subset, ranging from 0 to 1.

In contrast to importance, which focuses on individual token utility, we characterize the spatial distribution of the selected subset using the Hopkins Statistic (Hopkins and Skellam, 1954), a measure

that quantifies the degree of clustering in a dataset. A high Hopkins value indicates strong clustering, meaning that selected tokens concentrate in specific semantic regions and thus exhibit high redundancy.

Definition 3 (Diversity Metric via Hopkins Statistic)**.** *Let $\mathcal{S}$ denote the selected token subset with $|\mathcal{S}| = m$. We construct a reference set $\mathcal{R}$ by randomly sampling m points from the same feature space as $\mathcal{S}$. Let $d(x, \mathcal{Y})$ denote the cosine distance from point x to its nearest neighbor in set $\mathcal{Y}$. The Hopkins Statistic is defined as:*

$$H(\mathcal{S}) = \frac{\sum_{r \in \mathcal{R}} d(r, \mathcal{S})}{\sum_{r \in \mathcal{R}} d(r, \mathcal{S}) + \sum_{v \in \mathcal{S}} d(v, \mathcal{S} \setminus \{v\})}$$

In this formulation, $\mathcal{S} \setminus \{v\}$ denotes the set difference, representing the subset $\mathcal{S}$ excluding the specific token v to ensure the distance is calculated against its nearest neighbor.

Intuitively, $H(\mathcal{S}) \to 1$ indicates high redundancy due to significant clustering, while $H(\mathcal{S}) \to 0$ signifies a regularly spaced distribution with maximal semantic diversity.

3.2 Simulation on Real Token Manifolds

To identify the optimal strategy for harmonizing importance and diversity, we conduct a systematic analysis to explore their interaction. Specifically, we employ real visual tokens extracted from the Vision Transformer of the Qwen2.5-VL-7B-Instruct model as feature vectors. Real features are essential as they preserve complex manifold structures—such as semantic clustering and sparsity—that synthetic data typically fails to capture.

For token importance, we adopt a randomized approach where the score for each token is sampled independently from a uniform distribution $\mathcal{U}(0, 1)$. This setup decouples the evaluation of selection strategies from the bias of any specific pre-trained importance scorer.

We evaluate five representative strategies that make different trade-offs between importance and diversity:

- **Greedy Importance:** Selects tokens with the highest importance scores, ignoring diversity.
- **Greedy Diversity:** Iteratively selects the token that maximizes distance to the current subset via Farthest Point Sampling (Resende et al., 2010), prioritizing diversity over importance.
- **Naive Hybrid:** A two-stage approach that first selects top-k tokens by importance, then applies Farthest Point Sampling within this subset.
- **Determinantal Point Processes (DPP):** Models diversity probabilistically via the determinant of a kernel matrix (Macchi, 1975).
- **Maximal Marginal Relevance (MMR):** A joint optimization framework that explicitly balances importance and redundancy. We provide the detailed formulation of this mechanism in Section 3.3.

3.3 The Maximal Marginal Relevance (MMR) Mechanism

Maximal Marginal Relevance (MMR) (Carbonell and Goldstein-Stewart, 1998) provides a framework for this joint optimization. Initially proposed for information retrieval, the core idea of MMR is that an ideal result set should balance two criteria: high relevance to the query and low redundancy among selected items.

Adapting this principle to visual token pruning, the algorithm iteratively selects the token v^* from the candidate set $\mathcal{V} \setminus \mathcal{S}$ that maximizes the following objective:

$$\begin{aligned} v^* = \arg\max_{v_i \in \mathcal{V} \setminus \mathcal{S}} \big[& \lambda \cdot \mathrm{Imp}(v_i) \\ & - (1 - \lambda) \cdot \max_{v_j \in \mathcal{S}} \mathrm{Sim}(v_i, v_j) \big] \end{aligned}$$

where $\mathcal{V}$ represents the set of all visual tokens, $\mathcal{S}$ denotes the currently selected subset, $\mathrm{Imp}(\cdot)$ represents the normalized importance score, $\mathrm{Sim}(\cdot, \cdot)$ measures the pairwise similarity between tokens, and λ is a hyperparameter balancing the two terms.

By subtracting the maximum similarity between the candidate and the current subset $\mathcal{S}$, the algorithm explicitly penalizes tokens that are semantically close to any already selected token, while prioritizing important tokens.

3.4 Comparative Analysis against Heuristic Baselines

We conducted the simulation on 200 randomly sampled images from the MMBench dataset (Yuan Liu, 2023) to systematically evaluate the efficacy of the proposed strategies.

Figure 3 illustrates the trade-off between importance retention and diversity for each strategy. The theoretical optimum resides in the top-left corner,

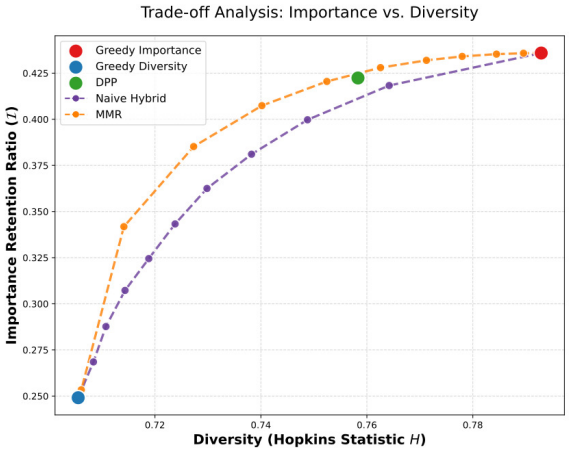

Figure 3: **Pareto Frontier Analysis.** We visualize the trade-off between the **Hopkins Statistic** (H) and the **Importance Retention Ratio** ($\mathcal{I}$). The ideal pruning strategy should approach the **top-left corner**, achieving a high Importance Retention Ratio ($\mathcal{I} \to 1$) while minimizing the Hopkins Statistic ($H \to 0$). The MMR mechanism (Orange) constructs a superior Pareto frontier that strictly dominates the Naive Hybrid strategy (Purple) and envelopes the DPP solution (Green).

corresponding to subsets that maximize $\mathcal{I}$ while minimizing H, thereby maximizing diversity. As illustrated, the single-objective baselines occupy the sub-optimal extremes: Greedy Importance (Red node) achieves maximum $\mathcal{I}$ at the cost of a high Hopkins Statistic ($H \approx 1$), whereas Greedy Diversity (Blue node) minimizes H but suffers from a low Importance Retention Ratio.

Crucially, the trajectory generated by MMR (Orange curve) forms a superior Pareto Frontier. It strictly dominates the Naive Hybrid strategy (Purple curve), maintaining a higher $\mathcal{I}$ for any given level of H, confirming the efficacy of our joint optimization framework. Furthermore, it effectively envelopes the DPP solution (Green node), demonstrating that our joint optimization framework provides the most robust mechanism for harmonizing these conflicting objectives.

4 Harmonizing Importance and Diversity via MMR

4.1 Token Importance Estimation

The analysis in Section 3 has demonstrated that the MMR mechanism effectively harmonizes diversity and importance. However, applying this framework in practice requires a computable importance metric.

To this end, we adopt the importance estimation mechanism of VisionSelector (Zhu et al., 2025a), which currently represents the state-of-the-art among importance-based pruning approaches. Specifically, this method employs a trainable estimation module coupled with a differentiable selection mechanism, *DiffTopK*, to learn token importance through end-to-end training. To maintain consistency with the training phase, we utilize the output of the *DiffTopK* mechanism as our raw importance scores, denoted as $\mathbf{w}$.

However, since MMR involves a direct subtraction between importance and similarity, both metrics must have comparable scales to prevent one from dominating the selection process. We therefore apply min-max normalization to the raw importance vector $\mathbf{w}$ to define the normalized importance metric:

$$\mathrm{Imp}(v_i) = \frac{w_i - \min(\mathbf{w})}{\max(\mathbf{w}) - \min(\mathbf{w}) + \epsilon}$$

where ϵ is a small constant for numerical stability. This procedure maps importance scores to the interval $[0, 1]$, ensuring they are commensurate with the similarity constraint.

4.2 Quantifying Collective Redundancy

In addition to importance, the MMR framework requires a metric to quantify semantic redundancy. In the latent feature space of MLLMs, tokens representing similar visual concepts tend to cluster together. Thus, we define the pairwise similarity between a candidate token v_i and a reference token v_j using cosine similarity:

$$\mathrm{Sim}(v_i, v_j) = \frac{v_i^\top v_j}{\|v_i\| \|v_j\|}$$

where $\|\cdot\|$ denotes the Euclidean norm. This metric enables the algorithm to identify tokens that are semantically similar to those already selected.

4.3 IDPruner: An MMR-based Selection Strategy

Building upon the normalized importance and semantic similarity metrics defined above, we formally present the **I**mportance and **D**iversity Pruner (**IDPruner**). This method harmonizes the two conflicting objectives within the MMR framework to iteratively construct the optimal subset. At each step t, IDPruner selects the token v^* from the remaining candidates $\mathcal{V} \setminus \mathcal{S}_{t-1}$ by maximizing the following objective:

$$v^* = \arg \max_{v_i \in \mathcal{V} \setminus \mathcal{S}_{t-1}} [\lambda \cdot \mathrm{Imp}(v_i) - (1 - \lambda) \cdot m_i]$$

Method	AI2D *EM*	ChartQA *Relaxed*	DocVQA *Anls*	MMB^{CN} *Score*	MMB *Score*	MME *Score*	MMStar *Avg*	OCRBench *Acc*	POPE *Acc*	SQA *EM*	VQA^{Text} *EM*	Avg
Baseline	82.48	83.68	94.90	80.41	83.08	1702	61.88	85.30	87.80	88.45	82.74	100.0%
Retain 25% Tokens (75% Compression Ratio)												
Importance-based methods												
FastV	75.68	68.20	81.20	73.20	76.12	1636	51.08	43.00	85.20	83.49	80.06	87.16%
VisionZip	77.40	67.20	71.48	76.12	78.78	1637	54.86	46.50	85.76	83.99	76.21	87.55%
HiPrune	77.49	68.60	73.52	76.03	78.09	1619	54.43	47.10	86.02	84.18	76.43	87.80%
VisionSelector	79.60	72.00	**93.24**	75.86	78.78	1688	55.78	72.50	86.74	85.08	80.39	94.22%
Diversity-based methods												
DivPrune	77.98	62.00	85.32	75.77	77.84	1650	52.97	58.40	85.88	83.94	75.88	89.26%
DART	74.35	60.80	78.90	73.88	76.72	1625	52.90	46.00	84.34	84.33	71.68	85.74%
Hybrid strategies												
VisPruner	77.62	68.04	77.39	75.69	78.87	1657	54.01	48.70	85.68	84.18	75.17	88.31%
SCOPE	78.92	71.20	85.40	**77.75**	79.38	1684	**56.86**	61.70	86.78	85.23	79.66	92.51%
IDPruner	**80.51**	**74.32**	93.16	76.63	**79.73**	**1695**	56.49	**74.00**	**87.06**	**85.52**	**80.83**	**95.18%**
Retain 10% Tokens (90% Compression Ratio)												
Importance-based methods												
FastV	67.23	39.48	51.90	53.26	55.58	1332	38.02	24.10	76.31	79.28	72.59	68.07%
VisionZip	70.60	41.56	37.94	66.67	71.05	1462	45.19	23.40	81.06	**83.24**	61.06	71.84%
HiPrune	69.82	43.96	39.89	67.44	70.88	1438	45.04	23.70	80.70	82.65	62.51	72.22%
VisionSelector	74.81	**62.68**	**87.00**	68.99	71.65	1569	46.93	**55.50**	82.69	81.95	**74.52**	85.39%
Diversity-based methods												
DivPrune	70.11	41.36	66.20	69.42	72.16	1529	44.46	31.80	81.91	80.96	62.72	76.09%
DART	67.88	34.84	49.86	63.92	67.35	1451	42.93	24.30	79.70	80.96	54.06	69.80%
Hybrid strategies												
VisPruner	69.88	42.68	50.85	66.84	70.96	1442	44.14	24.40	81.03	81.11	59.66	72.60%
SCOPE	71.63	50.04	56.45	71.22	**75.43**	1608	**48.74**	34.10	84.10	82.25	70.61	79.35%
IDPruner	**75.16**	62.48	85.98	**71.65**	74.66	**1618**	47.48	53.90	**85.43**	82.80	74.43	**86.47%**

Table 1: Comparison results on comprehensive Image-Language benchmarks on **Qwen-2.5-7B-Instruct**.

where $m_i = \max_{v_j \in \mathcal{S}_{t-1}} \text{Sim}(v_i, v_j)$ represents the maximum similarity between the candidate v_i and any token in the currently selected set, and $\lambda \in [0, 1]$ is the hyperparameter balancing importance and diversity.

Algorithm 1 IDPruner

Require: Tokens $\mathcal{V}$, Raw Importance Scores $\mathbf{w}$, Budget K, Hyperparameter λ
Ensure: Pruned subset $\mathcal{S}$

```
Imp ← (w − min w)/(max w − min w + ϵ)
S ← ∅, m ← fill(N, −1.0)
for t = 1 to K do
    if t = 1 then
        v* ← arg max_{v_i ∈ V} Imp(v_i)
    else
        v* ← arg max_{v_i ∉ S} [λImp(v_i) − (1 − λ)m_i]
    end if
    S ← S ∪ {v*}
    m ← max(m, Sim(V, v*))
end for
return S
```

To minimize computational overhead, we adopt an efficient updating strategy. Instead of recomputing the similarity scores for all pairs at every step, we maintain a vector $\mathbf{m} \in \mathbb{R}^N$ that tracks the maximum similarity for each candidate. After selecting v^*, we simply update this vector: $m_i \leftarrow \max(m_i, \text{Sim}(v_i, v^*))$. This implementation reduces the computational complexity from $O(K^2N)$ to $O(KN)$, rendering the overhead negligible relative to the model's forward pass. The complete procedure is summarized in Algorithm 1.

5 Experiments

5.1 Experimental Setup

Model Architectures. We conduct our main experiments on widely adopted MLLMs, including Qwen2.5-VL-7B-Instruct (Bai et al., 2025) and LLaVA-1.5-7B. (Liu et al., 2023b).

Evaluation benchmarks. We conduct comprehensive evaluations on image and video understanding tasks. For image-language understanding, we employ 10 widely-used datasets: MME (Fu et al., 2023), MMBench (Yuan Liu, 2023), MMStar (Chen et al., 2024b), POPE (Li et al., 2023), ScienceQA (Lu et al., 2022), AI2D (Kembhavi et al., 2016), TextVQA (Singh et al., 2019), ChartQA (Masry et al., 2022), DocVQA (Mathew et al., 2020), and OCRBench (Liu et al., 2024b). For video-language understanding, we include 3 benchmarks: Vinoground (Zhang et al., 2024a),

Method	AI2D *EM*	ChartQA *Relaxed*	DocVQA *Anls*	MMB^{CN} *Score*	MMB *Score*	MME *Score*	MMStar *Avg*	OCRBench *Acc*	POPE *Acc*	SQA *EM*	VQA^{Text} *EM*	Avg
Baseline	52.78	18.12	24.09	50.17	62.20	1463	32.74	19.80	85.86	66.19	47.78	100.0%
Retain 128 Tokens (77% Compression Ratio)												
Importance-based methods												
FastV	50.58	13.84	12.69	46.13	57.90	1213	30.89	12.10	72.04	65.44	31.76	81.59%
VisionZip	50.81	16.80	19.93	48.88	60.05	1374	32.52	18.50	82.28	66.34	45.40	94.86%
HiPrune	51.98	17.00	20.79	**49.74**	60.14	1386	32.20	18.50	82.14	66.19	45.54	95.63%
VisionSelector	50.74	16.20	20.83	49.23	60.31	1379	**34.22**	18.40	83.06	66.39	45.29	95.51%
Diversity-based methods												
DivPrune	51.65	16.36	18.58	46.13	57.99	1354	32.79	17.70	**85.16**	66.63	43.75	93.09%
DART	**52.82**	15.32	15.02	44.50	56.53	1309	30.55	14.00	77.08	66.04	34.90	85.69%
Hybrid strategies												
VisPruner	51.68	16.56	20.05	**49.74**	**60.57**	1382	32.47	18.20	83.57	66.63	46.21	95.39%
SCOPE	51.30	17.20	**21.56**	49.14	60.22	1374	32.68	**19.30**	84.41	66.58	**46.84**	96.77%
IDPruner	51.55	**17.40**	21.55	49.23	60.40	**1428**	33.44	18.80	84.69	**66.88**	46.39	**97.26%**
Retain 64 Tokens (88% Compression Ratio)												
Importance-based methods												
FastV	49.42	12.24	9.74	40.38	47.08	964	27.78	4.00	61.38	63.71	17.69	66.68%
VisionZip	51.20	15.76	15.75	46.31	56.70	1289	31.40	16.40	78.18	66.44	42.75	89.14%
HiPrune	51.20	15.68	16.18	46.13	57.30	1257	31.15	17.40	78.17	66.48	43.00	89.56%
VisionSelector	50.65	15.12	17.29	46.65	**58.33**	1310	**32.58**	17.00	79.94	65.84	42.98	90.49%
Diversity-based methods												
DivPrune	50.00	15.12	15.69	44.16	54.64	1271	31.68	15.70	**84.18**	66.44	40.83	87.82%
DART	**51.55**	15.20	12.35	40.29	53.26	1195	28.50	12.40	71.03	**66.83**	30.98	79.88%
Hybrid strategies												
VisPruner	51.33	15.28	16.29	44.76	57.39	1315	31.02	17.00	81.20	66.73	43.45	89.77%
SCOPE	50.71	15.56	17.84	46.82	58.16	1320	32.18	**17.70**	83.07	66.63	44.33	91.90%
IDPruner	50.55	**16.32**	**18.59**	**46.99**	57.65	**1329**	31.49	17.30	83.83	66.68	**44.73**	**92.34%**
Retain 32 Tokens (94% Compression Ratio)												
Importance-based methods												
FastV	49.03	11.96	8.38	30.24	38.23	844	27.15	3.00	57.50	64.20	10.98	59.83%
VisionZip	51.17	14.04	13.10	39.35	50.43	1133	29.90	13.90	72.42	66.98	38.15	81.15%
HiPrune	51.42	14.32	13.22	39.95	51.98	1149	29.28	14.20	73.26	66.78	38.29	81.87%
VisionSelector	50.06	13.52	14.00	43.81	54.90	1194	**32.07**	14.80	75.04	65.49	38.94	84.12%
Diversity-based methods												
DivPrune	50.74	14.20	12.25	37.89	51.72	1149	28.68	13.80	80.47	65.69	36.21	80.78%
DART	51.10	13.32	10.79	33.51	47.08	1069	29.17	10.70	63.86	**67.58**	25.35	73.03%
Hybrid strategies												
VisPruner	**51.59**	14.28	12.95	37.97	50.95	1165	29.40	14.00	75.14	67.23	37.75	81.47%
SCOPE	51.17	**15.00**	14.91	43.64	**55.58**	1239	31.36	15.10	79.50	66.83	40.14	86.57%
IDPruner	51.23	14.80	**15.11**	**44.76**	55.50	**1254**	30.58	**15.90**	**81.43**	66.73	**41.46**	**87.43%**

Table 2: Comparison results on comprehensive Image-Language benchmarks on **LLaVA-1.5-7B**.

VideoMME (Fu et al., 2025), and SEED-Bench (Li et al., 2024b). To ensure fair comparison and reproducibility, we utilize the LMMs-Eval framework (Zhang et al., 2024b), strictly following the default settings and metrics for each task.

Comparison methods. We compare IDPruner with representative state-of-the-art approaches across different paradigms, including importance-based methods like FastV (Chen et al., 2024a), VisionZip (Yang et al., 2025b), HiPrune (Liu et al., 2025), and VisionSelector (Zhu et al., 2025a), diversity-based methods like DivPrune (Alvar et al., 2025) and DART (Wen et al., 2025), as well as hybrid strategies that combine multiple criteria, such as VisPruner (Zhang et al., 2024c), and SCOPE (Deng et al., 2025).

Implementation Details. Unless otherwise specified, the hyperparameter λ of IDPruner, which balances importance and diversity, is set to 0.5.

5.2 Main Results

Results on Qwen2.5-VL-7B-Instruct. We evaluate our method on Qwen2.5-VL-7B-Instruct under 25% and 10% token retention settings. As shown in Table 1, IDPruner achieves state-of-the-art average scores of 95.18% and 86.47%, respectively. Compared to existing strategies, our method achieves a better balance between keeping fine details and maintaining global context. Specifically, for tasks requiring fine details, such as OCRBench, our method ranks among the top two, while also maintaining global information to surpass VisionSelector on hallucination benchmarks, including POPE. Consequently, on benchmarks such as MME and AI2D, which require both overall understanding and detailed capture, IDPruner demonstrates a clear

lead over other methods.

Results on LLaVA-1.5-7B. We extend our experiments to the LLaVA-1.5-7B model, which operates with a fixed resolution of 576 visual tokens per image. Accordingly, we evaluate performance under three distinct retention settings: 128, 64, and the extreme 32 tokens. As shown in Table 2, IDPruner consistently achieves state-of-the-art results across all pruning ratios. While VisionSelector is surpassed by the hybrid method SCOPE on this architecture, IDPruner maintains its lead, achieving an average score of 87.43% even with only 32 tokens, demonstrating its robustness across diverse architectures.

In summary, our method exhibits remarkable performance consistency across a diverse range of architectures. Notably, strong baselines exhibit architecture-specific vulnerabilities; for instance, VisionSelector underperforms on LLaVA-1.5, whereas SCOPE loses competitiveness on the advanced LLaVA-OneVision-1.5, as detailed in Appendix A. In contrast, IDPruner maintains exceptional robustness. It consistently achieves state-of-the-art results across all evaluated models, validating the universality of our framework in harmonizing token importance and diversity.

Method	Vinoground *Group*	VideoMME *Perception*	SEED-Bench *All*	Avg
Baseline	20.20	61.33	74.12	100.0%
FastV	12.80	59.44	69.22	84.56%
VisionZip	12.80	59.67	72.11	85.98%
HiPrune	11.80	59.93	71.97	84.41%
VisionSelector	10.80	59.19	70.75	81.81%
DivPrune	**14.00**	58.00	72.11	87.06%
DART	12.60	59.52	71.27	85.19%
VisPruner	11.40	59.44	71.89	83.45%
SCOPE	12.80	**60.00**	72.63	86.40%
IDPruner	13.40	59.48	**72.68**	**87.13%**

Table 3: Comparison results on Video-Language benchmarks on **Qwen2.5-VL-7B-Instruct** with **25% token retention**.

5.3 IDPruner for Video Understanding

Beyond static image benchmarks, we extend IDPruner to video understanding tasks, evaluating its performance on Vinoground, VideoMME, and SEED-Bench at a 75% pruning ratio. As shown in Table 3, purely importance-based methods exhibit significant performance degradation. This is primarily due to their inability to handle the high temporal redundancy in videos. In contrast, diversity-based methods maintain strong performance with an average score of 87.06%. Notably, IDPruner achieves the best average performance of 87.13% by jointly considering both the preservation of important details and the reduction of temporal redundancy.

Method	FA	Prefill(ms)	E2E Latency(ms)
VisPruner	×	1459.95	1600.81
SCOPE	×	1677.81	1818.40
IDPruner	✓	**1337.76**	**1478.32**

Table 4: Efficiency analysis on Vinoground on **Qwen2.5-VL-7B-Instruct** with **25% token retention**. **FA**: FlashAttention compatibility.

5.4 Efficiency and Practicality

We compare the efficiency of hybrid pruning strategies on Qwen2.5-VL-7B using the Vinoground benchmark at a 75% pruning ratio. As shown in Table 4, IDPruner achieves the best efficiency among hybrid strategies, due to its lightweight diversity calculation and being attention-map-free. This design ensures full compatibility with FlashAttention, yielding the lowest prefill time of 1337.76 ms and an end-to-end latency of 1478.32 ms.

6 Conclusion

Recent progress in visual token pruning shows that hybrid strategies are surpassing methods that rely only on importance or diversity, becoming the new standard in this field. However, there is a lack of systematic analysis on how to effectively harmonize these two objectives. In this study, we provide a framework to analyze this trade-off and demonstrate that the Maximal Marginal Relevance (MMR) mechanism is an effective strategy to achieve an optimal balance. Based on this insight, we propose IDPruner, a method that explicitly balances token importance and semantic redundancy. Extensive evaluations show that our method achieves state-of-the-art performance and remains robust across different model architectures. We believe this work offers a solid foundation for systematically balancing importance and diversity, enabling more efficient MLLMs.

References

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. 2025. Divprune: Diversity-based visual token pruning for large multimodal models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 9392–9401.

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

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, and 1 others. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Jaime G. Carbonell and Jade Goldstein-Stewart. 1998. The use of mmr, diversity-based reranking for reordering documents and producing summaries. *ACM SIGIR Forum*, 51:209 – 210.

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

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and 1 others. 2024b. Are we on the right way for evaluating large vision-language models? *arXiv preprint arXiv:2403.20330*.

Tri Dao, Daniel Y. Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. FlashAttention: Fast and memory-efficient exact attention with IO-awareness. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Jinhong Deng, Wen Li, Joey Tianyi Zhou, and Yang He. 2025. Scope: Saliency-coverage oriented token pruning for efficient multimodel llms. *ArXiv*, abs/2510.24214.

Mohamed Achraf Dhouib, Davide Buscaldi, Sonia Vanier, and Aymen Shabou. 2025. Pact: Pruning and clustering-based token reduction for faster visual language models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 14582–14592.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *ArXiv*, abs/2306.13394.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Jiayi Han, Liang Du, Yiwen Wu, Guanming Liang, Xiangguo Zhou, Weibo Zheng, Donghong Han, and Zixun Sun. 2025. Adav: Adaptive text-visual redirection for vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4985–4997.

Brian Hopkins and J. Gordon Skellam. 1954. A new method for determining the type of distribution of plant individuals. *Annals of Botany*, 18:213–227.

Aniruddha Kembhavi, Michael Salvato, Eric Kolve, Minjoon Seo, Hannaneh Hajishirzi, and Ali Farhadi. 2016. A diagram is worth a dozen images. *ArXiv*, abs/1603.07396.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles*.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *ArXiv*, abs/2408.03326.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2024b. Seedbench: Benchmarking multimodal large language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13299–13308.

Yangfu Li, Hongjian Zhan, Tianyi Chen, Qi Liu, and Yue Lu. 2025. Why 1 + 1 < 1 in visual token pruning: Beyond naive integration via multi-objective balanced covering. *ArXiv*, abs/2505.10118.

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

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning.

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

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

Jizhihui Liu, Feiyi Du, Guangdao Zhu, Niu Lian, Jun Li, and Bin Chen. 2025. Hiprune: Training-free visual token pruning via hierarchical attention in vision-language models. *arXiv preprint arXiv:2508.00553*.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024b. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12).

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*.

Odile Macchi. 1975. The coincidence approach to stochastic point processes. *Advances in Applied Probability*, 7(1):83–122.

Ahmed Masry, Do Xuan Long, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. *arXiv preprint arXiv:2203.10244*.

Minesh Mathew, Dimosthenis Karatzas, R Manmatha, and CV Jawahar. 2020. Docvqa: A dataset for vqa on document images. corr abs/2007.00398 (2020). *arXiv preprint arXiv:2007.00398*.

Mauricio GC Resende, Rafael Martí, Micael Gallego, and Abraham Duarte. 2010. Grasp and path relinking for the max–min diversity problem. *Computers & Operations Research*, 37(3):498–508.

Kele Shao, Keda Tao, Kejia Zhang, Sicheng Feng, Mu Cai, Yuzhang Shang, Haoxuan You, Can Qin, Yang Sui, and Huan Wang. 2025. When tokens talk too much: A survey of multimodal long-context token compression across images, videos, and audios. *ArXiv*, abs/2507.20198.

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

Jintao Tong, Wenwei Jin, Pengda Qin, Anqi Li, Yixiong Zou, Yuhong Li, Yuhua Li, and Ruixuan Li. 2025. Flowcut: Rethinking redundancy via information flow for efficient vision-language models. *ArXiv*, abs/2505.19536.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, and 1 others. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*.

Zekun Wang, Minghua Ma, Zexin Wang, Rongchuan Mu, Liping Shan, Ming Liu, and Bing Qin. 2025. Effivlm-bench: A comprehensive benchmark for evaluating training-free acceleration in large vision-language models. *Preprint*, arXiv:2506.00479.

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

Long Xing, Qidong Huang, Xiaoyi Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and 1 others. 2024. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*.

Cheng Yang, Yang Sui, Jinqi Xiao, Lingyi Huang, Yu Gong, Chendi Li, Jinghua Yan, Yu Bai, Ponnuswamy Sadayappan, Xia Hu, and 1 others. 2025a. 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*, pages 19803–19813.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. 2025b. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19792–19802.

Xubing Ye, Yukang Gan, Yixiao Ge, Xiao-Ping Zhang, and Yansong Tang. 2025. Atp-llava: Adaptive token pruning for large vision language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24972–24982.

Yuanhan Zhang Bo Li Songyang Zhang Wangbo Zhao Yike Yuan Jiaqi Wang Conghui He Ziwei Liu Kai Chen Dahua Lin Yuan Liu, Haodong Duan. 2023. Mmbench: Is your multi-modal model an all-around player? *arXiv:2307.06281*.

Jianrui Zhang, Cai Mu, and Yong Jae Lee. 2024a. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *arXiv*.

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

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. 2024c. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms.

Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. 2024d. [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster. *arXiv e-prints*, pages arXiv–2412.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. 2025a. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *ArXiv*, abs/2506.10967.

Weichen Zhang, Zhui Zhu, Ningbo Li, Kebin Liu, and Yunhao Liu. 2025b. Adaptinfer: Adaptive token pruning for vision-language model inference with dynamical text guidance. *arXiv preprint arXiv:2508.06084*.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and 1 others. 2024e. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*.

Zixuan Zhou, Xuefei Ning, Ke Hong, Tianyu Fu, Jiaming Xu, Shiyao Li, Yuming Lou, Luning Wang, Zhihang Yuan, Xiuhong Li, Shengen Yan, Guohao Dai, Xiao-Ping Zhang, Yuhan Dong, and Yu Wang. 2024. A survey on efficient inference for large language models. *ArXiv*, abs/2404.14294.

Jiaying Zhu, Yurui Zhu, Xin Lu, Wenrui Yan, Dong Li, Kunlin Liu, Xueyang Fu, and Zheng-Jun Zha. 2025a. Visionselector: End-to-end learnable visual token compression for efficient multimodal llms. *Preprint*, arXiv:2510.16598.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025b. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Xin Zou, Di Lu, Yizhou Wang, Yibo Yan, Yuanhuiyi Lyu, Xu Zheng, Linfeng Zhang, and Xuming Hu. 2025. Don't just chase "highlighted tokens" in mllms: Revisiting visual holistic context retention. *ArXiv*, abs/2510.02912.

A Additional Experimental Results

A.1 Results on Qwen2.5-VL-3B-Instruct

To evaluate the scalability of our method on smaller language models, we conduct experiments on **Qwen2.5-VL-3B-Instruct**. As shown in Table 5, IDPruner consistently outperforms competitive baselines at both 25% and 10% token retention ratios. Notably, when retaining 25% of the tokens, our method achieves an average score of 94.42%, effectively matching the unpruned baseline. Even under the aggressive 10% retention setting, IDPruner maintains a high average performance of 85.71%, outperforming the second-best method (VisionSelector) by 1.29%.

A.2 Results on LLaVA-OneVision-1.5-8B-Instruct

We further assess the cross-architecture generalization on **LLaVA-OneVision-1.5-8B-Instruct**, which integrates advanced visual encoding strategies. As shown in Table 6, IDPruner achieves the best results among existing state-of-the-art methods. Under the 25% retention setting, our method achieves an average score of 92.00%, outperforming the strongest baseline, VisionSelector, by 0.37%. In the more challenging 10% retention scenario, IDPruner exhibits strong robustness, achieving an average score of 81.55%. It significantly outperforms purely importance-based methods such as VisionZip and HiPrune, which suffer from severe degradation due to the loss of global context. Additionally, it surpasses the competitive VisionSelector by 1.44%, confirming that harmonizing importance and diversity is particularly effective for advanced architectures.

B Ablation Study: Integration Strategies and Hyperparameters

We investigate the efficacy of different integration strategies and the impact of the hyperparameter λ, which controls the trade-off between token importance and diversity. Using VisionSelector as the fixed base importance estimator, we compare our **IDPruner (MMR)** mechanism against two representative baselines: a determinantal point process based method (**DPP**) and a **Naive Hybrid** strategy that combines importance filtering with Farthest Point Sampling (FPS). Table 7 summarizes the results on **Qwen2.5-VL-7B-Instruct** at a 25% token retention ratio.

Superiority of MMR Mechanism. The integration strategy plays a pivotal role in model performance. As evidenced in Table 7, **IDPruner** consistently outperforms the Naive Hybrid strategy across comparable λ settings and also surpasses the DPP-based baseline. The Naive Hybrid approach typically prioritizes tokens with the highest importance scores before applying Farthest Point Sampling (FPS) to enhance diversity. However, this two-stage paradigm fails to address the inherent redundancy among high-importance tokens, resulting in a selected subset that lacks sufficient diversity. In contrast, IDPruner employs a unified scoring mechanism that simultaneously manages importance and redundancy. By dynamically penalizing semantically repetitive tokens during selection, our method achieves a more effective balance, thereby demonstrating superior robustness over heuristic hybrid strategies.

Hyperparameter Selection. The hyperparameter λ controls the balance between token importance and semantic diversity. For IDPruner, the performance follows an inverted U-shape pattern, peaking at $\lambda = 0.5$ with an average performance of **95.56%**. This confirms that setting $\lambda = 0.5$ successfully strikes an optimal balance between token importance and semantic diversity, enabling IDPruner to leverage both properties for maximum performance.

C Limitations

Despite the promising results achieved by IDPruner, we acknowledge certain limitations in this study. First, constrained by computational resources, we have not yet evaluated our method on long-context video understanding benchmarks. This restricts the comprehensive verification of our method's effectiveness in scenarios involving extremely long temporal sequences, thereby limiting the scope of applicable scenarios. Second, due to time constraints, we did not conduct a fine-grained measurement or exhaustive search for the hyperparameter λ. While the current settings demonstrate strong robustness, a more thorough optimization could potentially yield further performance improvements.

D Visualization

To intuitively understand how IDPruner harmonizes importance and diversity compared to existing approaches, we visualize the spatial distri-

Method	AI2D *EM*	ChartQA *Relaxed*	DocVQA *Anls*	MMB^{CN} *Score*	MMB *Score*	MME *Score*	MMStar *Avg*	OCRBench *Acc*	POPE *Acc*	SQA *EM*	VQA^{Text} *EM*	Avg
Baseline	79.11	83.56	92.48	73.28	77.32	1517	56.05	80.10	87.41	80.81	78.79	100.0%
Retain 25% Tokens (75% Compression Ratio)												
Importance-based methods												
FastV	72.70	70.04	75.98	63.40	66.92	1437	47.39	36.60	86.42	79.33	73.51	86.02%
VisionZip	74.19	71.32	70.11	67.35	71.22	1452	49.37	42.50	85.51	81.36	68.12	87.34%
HiPrune	73.83	72.76	72.10	67.27	72.34	1449	48.93	41.30	85.86	80.91	69.27	87.67%
VisionSelector	75.19	73.72	**90.24**	68.81	72.59	**1521**	49.97	61.80	85.36	80.37	76.86	93.62%
Diversity-based methods												
DivPrune	73.06	62.96	78.46	67.10	71.82	1459	48.38	51.40	**86.81**	80.22	68.91	88.15%
DART	71.08	65.20	79.72	65.38	71.05	1428	48.78	41.80	80.97	80.91	68.25	86.17%
Hybrid strategies												
VisPruner	74.29	68.20	72.52	67.35	70.88	1458	49.74	44.80	86.59	**81.46**	69.62	87.87%
SCOPE	75.84	74.00	82.40	68.81	72.94	1471	**50.35**	56.00	86.62	80.96	74.04	91.98%
IDPruner	**75.94**	**75.84**	90.00	**69.42**	**73.80**	1505	49.49	**64.90**	86.26	80.42	**76.90**	**94.42%**
Retain 10% Tokens (90% Compression Ratio)												
Importance-based methods												
FastV	65.87	29.72	36.89	48.37	51.98	1257	37.28	13.90	79.50	77.05	57.75	65.30%
VisionZip	67.65	51.60	37.88	59.62	63.06	1338	42.82	21.40	81.14	80.47	51.56	72.75%
HiPrune	67.75	53.20	41.15	59.45	63.14	1326	41.08	20.30	80.90	**80.96**	53.31	73.00%
VisionSelector	70.50	**65.92**	**79.94**	59.97	64.69	1374	42.86	45.20	82.66	80.61	**71.57**	84.42%
Diversity-based methods												
DivPrune	67.71	43.12	58.03	61.25	65.12	1389	40.43	27.90	82.24	79.18	56.87	75.50%
DART	67.49	47.56	60.23	57.99	63.83	1299	42.18	23.40	74.20	78.63	58.02	74.09%
Hybrid strategies												
VisPruner	67.75	47.92	48.65	59.28	63.32	1305	41.51	22.50	78.74	79.77	54.95	73.19%
SCOPE	69.75	56.24	55.01	**64.26**	67.18	1390	**44.35**	30.80	83.34	80.47	62.58	79.37%
IDPruner	**71.79**	63.32	79.38	63.57	**68.21**	**1438**	44.05	**45.50**	**84.51**	80.57	70.02	**85.71%**

Table 5: Comparison results with different methods on **Qwen2.5-VL-3B-Instruct**.

Method	AI2D *EM*	ChartQA *Relaxed*	DocVQA *Anls*	MMB^{CN} *Score*	MMB *Score*	MME *Score*	MMStar *Avg*	OCRBench *Acc*	POPE *Acc*	SQA *EM*	VQA^{Text} *EM*	Avg
Baseline	84.20	86.40	97.87	78.52	85.31	1594	68.25	80.90	88.91	98.76	79.65	100.0%
Retain 25% Tokens (75% Compression Ratio)												
Importance-based methods												
FastV	74.48	39.16	56.53	70.36	77.66	1440	50.21	33.60	81.53	89.19	52.54	75.05%
VisionZip	69.95	27.48	23.69	65.89	75.26	1419	47.01	20.80	80.46	84.78	36.15	65.14%
HiPrune	70.27	19.76	21.01	63.57	72.51	1339	46.58	19.60	77.29	81.36	30.21	61.59%
VisionSelector	77.85	**76.32**	**94.08**	**74.74**	80.07	1569	57.42	55.50	86.73	94.70	**77.53**	91.63%
Diversity-based methods												
DivPrune	78.85	65.52	81.53	73.20	80.67	1533	56.68	51.50	**88.14**	91.42	75.00	88.12%
DART	77.82	69.00	84.77	72.25	79.81	1564	56.80	46.90	84.77	93.11	73.89	87.83%
Hybrid strategies												
VisPruner	75.45	45.88	61.00	69.93	76.55	1468	50.68	36.70	85.82	89.24	69.01	79.01%
SCOPE	78.21	54.84	75.68	70.27	78.18	1505	54.52	46.60	87.02	90.18	72.77	84.30%
IDPruner	**79.18**	74.48	91.82	73.71	**81.27**	**1588**	**57.97**	**57.80**	88.13	**95.14**	77.50	**92.00%**
Retain 10% Tokens (90% Compression Ratio)												
Importance-based methods												
FastV	70.95	22.72	29.77	62.46	69.07	1303	43.66	16.70	73.83	82.35	38.71	62.08%
VisionZip	69.14	19.64	13.20	58.16	65.38	1259	42.12	12.00	76.59	78.53	21.28	56.09%
HiPrune	68.39	16.24	12.86	56.01	63.83	1208	42.13	9.10	73.72	77.29	18.97	53.92%
VisionSelector	73.38	**61.00**	**71.31**	68.04	74.91	1466	**50.74**	34.60	82.62	88.05	67.67	80.11%
Diversity-based methods												
DivPrune	73.74	37.68	56.57	67.18	73.80	1477	46.31	30.10	**85.33**	86.51	65.98	75.02%
DART	72.80	40.48	52.77	63.14	70.19	1378	47.42	27.30	75.79	84.88	61.58	71.65%
Hybrid strategies												
VisPruner	69.82	25.20	35.18	61.77	67.70	1375	43.24	18.20	78.33	84.23	54.53	65.46%
SCOPE	72.05	33.04	47.65	66.07	73.54	1471	48.25	25.70	84.11	85.97	61.02	72.35%
IDPruner	**74.45**	57.92	70.43	**69.67**	**75.69**	**1511**	50.02	**38.70**	85.02	**88.70**	**72.37**	**81.55%**

Table 6: Comparison results with different methods on **LLaVA-OneVision-1.5-8B-Instruct**.

Method	AI2D *EM*	ChartQA *Relaxed*	DocVQA *Anls*	MME *Score*	OCRBench *Acc*	POPE *Acc*	SQA *EM*	VQA^{Text} *EM*	Avg
Baseline	82.48	83.68	94.90	1701	85.30	87.80	88.45	82.74	100.00
Strategy 1: DPP + VisionSelector									
DPP+VisionSelector	79.70	73.36	93.02	1691	73.00	86.96	84.73	80.69	94.95
Strategy 2: IDPruner (MMR Mechanism)									
IDPruner ($\lambda = 0.1$)	78.08	67.36	87.94	1680	62.00	86.59	83.59	77.65	90.78
IDPruner ($\lambda = 0.3$)	80.05	73.88	91.26	1671	70.50	86.91	84.09	79.69	94.10
IDPruner ($\lambda = 0.5$)	**80.51**	74.32	93.16	1695	**74.00**	87.06	**85.52**	80.83	**95.56**
IDPruner ($\lambda = 0.7$)	80.25	74.12	93.35	**1710**	**74.00**	87.07	85.13	80.75	95.56
IDPruner ($\lambda = 0.9$)	79.66	72.72	93.29	1705	72.80	86.96	84.88	80.61	94.97
Strategy 3: Naive Hybrid Selector									
Hybrid ($\lambda = 0.1$)	78.79	66.52	86.90	1700	59.90	86.09	83.64	78.12	90.47
Hybrid ($\lambda = 0.3$)	79.50	72.72	90.71	1704	62.60	86.56	83.34	79.91	92.73
Hybrid ($\lambda = 0.5$)	79.95	74.36	92.20	1702	64.30	**87.17**	84.33	80.82	93.84
Hybrid ($\lambda = 0.7$)	79.18	**75.08**	93.10	1680	66.40	86.63	84.83	**80.94**	94.10
Hybrid ($\lambda = 0.9$)	79.31	73.84	**93.42**	1681	71.90	86.66	84.58	80.45	94.69

Table 7: Ablation study of integration strategies on Qwen2.5-VL-7B-Instruct with 25% token retention. We use VisionSelector as the base importance scorer. λ controls the trade-off between importance and diversity.

bution of retained visual tokens across multiple samples. Figure 4 presents a comparison of token selection masks under a 25% retention ratio. As consistently observed across diverse scenes, **DivPrune** tends to produce a uniform distribution, often overlooking semantic details. **VisionSelector** overly concentrates on foreground objects at the expense of background information coverage. In contrast, **IDPruner** successfully balances both, capturing salient features while maintaining essential background context necessary for global reasoning.

E Empirical Verification of Non-Negative Similarity

A potential concern regarding the MMR mechanism is the behavior of the redundancy penalty term, $(1 - \lambda) \cdot \text{Sim}(v_i, v_j)$. If the cosine similarity $\text{Sim}(v_i, v_j)$ were to yield negative values (implying an angle $\theta > 90^\circ$ between feature vectors), the intended penalty would transform into a reward.

To address this validity concern, we empirically analyzed the geometric properties of the visual token space. We randomly selected 100 images from the MMBench dataset and computed the pairwise angles between all visual tokens extracted from the **Qwen2.5-VL-7B-Instruct** model.

As illustrated in Figure 5, the distribution of pairwise angles exhibits a distinct pattern. The distribution is overwhelmingly concentrated within the range of $[0^\circ, 85^\circ]$, with a peak density at approximately 74°. Crucially, there is zero probability mass beyond the 90° threshold (indicated by the red dashed line).

Since $\text{Sim}(v_i, v_j) = \cos(\theta_{ij})$ and $\cos(\theta) \geq 0$ for all $\theta \in [0^\circ, 90^\circ]$, this empirical evidence confirms that all similarity scores in our framework are strictly non-negative. Consequently, the term $(1 - \lambda) \cdot \text{Sim}(v_i, v_j)$ consistently functions as a redundancy penalty, validating the theoretical soundness of our IDPruner formulation.

F Statement on the Use of AI Assistants

In accordance with the ACL submission policies, we hereby declare the use of AI assistants in the preparation of this manuscript. We utilized AI assistants for writing refinement, including grammar correction, vocabulary enhancement, and proofreading to improve readability. We emphasize that all scientific claims, experimental designs, core concepts, and logical arguments presented in this work are the original contributions of the authors. All AI-generated content was meticulously reviewed and verified by the authors to ensure accuracy and adherence to academic standards; the authors assume full responsibility for the content of this paper.

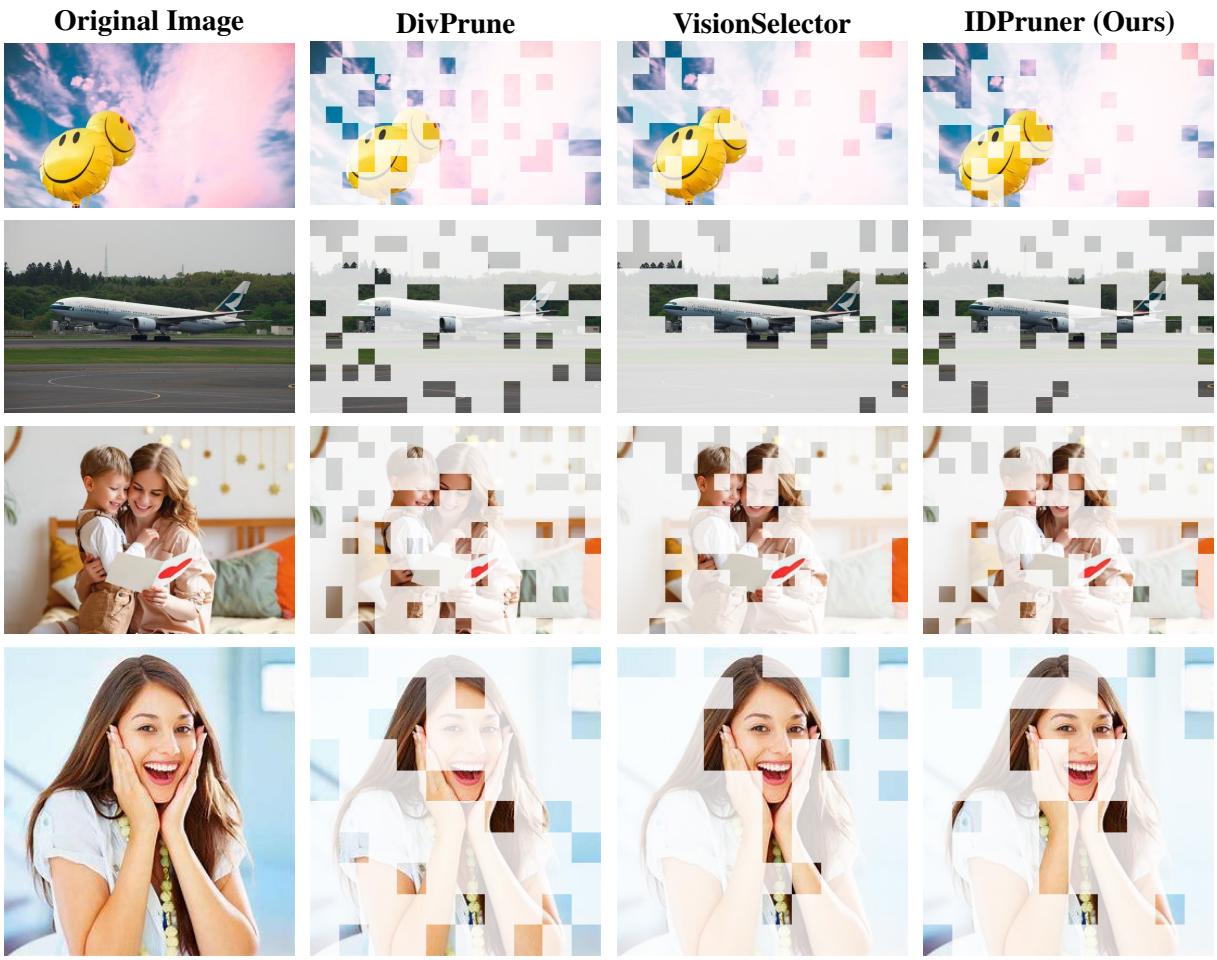

Figure 4: **Visualization of retained visual tokens across different samples from MMBench.** Columns from left to right: Original Image, DivPrune, VisionSelector, and IDPruner. **DivPrune** maintains global coverage but often neglects the semantic subject. **VisionSelector** clusters heavily on salient objects, resulting in redundancy and background loss. **IDPruner** achieves a superior balance, preserving intricate details of the subject while maintaining essential background context for global reasoning.

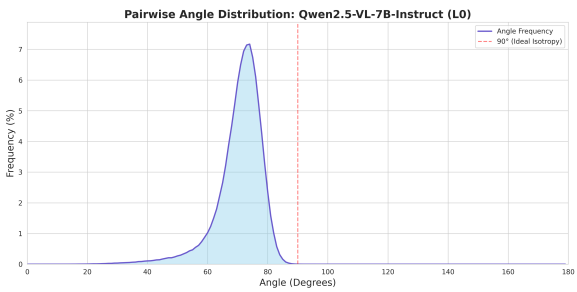

Figure 5: **Distribution of pairwise angles between visual tokens.** We calculated the angles for all token pairs across 100 images from MMBench using Qwen2.5-VL-7B. The distribution is entirely concentrated within the acute angle range ($< 90°$), peaking around $74°$. The absence of obtuse angles ($> 90°$, right of the red dashed line) guarantees that the cosine similarity metric remains strictly non-negative.