IFD-Prune: Importance-Fused Diverse Visual Token Pruning for Efficient Vision-Language Inference

Anonymous ACL submission

Abstract

Although Multimodal Large Language Models (MLLMs) excel in visual-language understanding, the quadratic complexity induced by massive visual tokens causes significant computational overhead. Existing visual token pruning strategies often rely on single-dimensional metrics, failing to balance image-intrinsic global context with text-guided relevance or effectively eliminate feature redundancy. To address this, we propose IFD-Prune, a training-free, Plug-and-Play visual token pruning framework. Specifically, we design a dual-criteria importance mechanism that explicitly fuses intrinsic visual saliency and cross-modal text relevance. Furthermore, we formulate visual token pruning as a maximum volumetric information problem, utilizing iterative greedy orthogonal projection to select tokens that span the largest effective hypervolume in the feature space. Extensive experiments demonstrate that IFD-Prune outperforms state-of-the-art methods. Notably, on LLaVA-1.5-7B, our method reduces visual tokens by 88.9% and FLOPs by 63.8% while robustly retaining 96.87% of the original performance, achieving a superior efficiency-accuracy trade-off.

1 Introduction

In recent years, the evolution of Large Language Models (LLMs) (Touvron et al., 2023; Achiam et al., 2023) has transcended the unimodal text domain, rapidly expanding into the multimodal realm. Multimodal Large Language Models(MLLMs) (Bai et al., 2023; Li et al., 2023a; Lin et al., 2024) achieve powerful cross-modal reasoning capabilities by encoding visual inputs into discrete sequences of visual tokens and aligning them with text tokens. However, this mechanism introduces significant computational overhead: taking LLaVA-v1.5 (Liu et al., 2023a) as an example, a single image is encoded into 576 tokens, while in LLaVA-Next (Liu et al., 2024), this number surges to 2880.

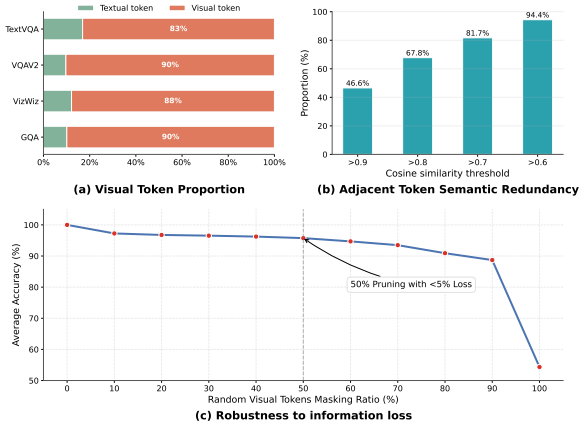


Figure 1: Motivation analysis of visual token redundancy in MLLMs. (a) **Token Distribution**: Proportion of visual tokens across different tasks. **Local Semantic Redundancy**: Percentage of tokens whose cosine similarity with at least one of their 8-neighbors exceeds a threshold τ. The high values indicate significant spatial-semantic overlap. (c)**Information Resilience**: Average performance degradation under random token masking across varying ratios. All the above experiments were conducted on the LLaVA-v1.5-7B model.

As shown in Figure 1 (a) , our statistics on LLaVA-v1.5-7B indicate that in general tasks, visual tokens often account for more than 80% of the total sequence length, and even exceed 90% in tasks like VQAv2 (Goyal et al., 2016). This **Vision-Dominant** input distribution not only drastically increases memory occupation but also becomes a major bottleneck for inference latency. This forces us to confront a core question: *To understand an image, are all visual tokens truly necessary?*

To answer this question, we conducted two pilot studies using LLaVA-v1.5-7B as the base model. First, we analyzed the local semantic redundancy of visual encoding. As shown in Figure 1 (b), the cosine similarity between adjacent visual tokens reveals that over 45% of adjacent token pairs exhibit similarity higher than 0.9, indicating high repetition in the feature space. Second, we tested the

sparsity of information density. As shown in Figure 1 (c), even with the most naive "Random Masking" strategy discarding 50% of visual tokens, the model can still maintain approximately 95% of its original performance. These experimental results strongly confirm that current MLLMs visual representations possess extremely high redundancy, and a large number of tokens carry only negligible minimal marginal information gain during inference. When the masking ratio is further increased to 80%, the performance of the random strategy collapses; therefore, how to identify visual tokens holding key information to maintain high model performance becomes a critical challenge.

Addressing this issue, recent research has proposed various visual token pruning and merging strategies (Bolya et al., 2022; Xing et al., 2024; Liu et al., 2025). Approaches like FastV (Chen et al., 2024) and SparseVLM (Zhang et al., 2024b) rely on Text-Vision Attention to assess token importance, retaining regions most relevant to text prompts. Others, such as VisionZip(Yang et al., 2024b), utilize the CLS token attention within the vision encoder to select visual token with high information content, while methods like DivPrune (Alvar et al., 2025) focus on maintaining visual token diversity by maximizing feature distances. However, these existing methods still suffer from significant limitations: attention-based methods often overlook spatial redundancy, resulting in the retention of highly similar tokens; conversely, diversity-oriented methods often ignore the dynamic guidance of user instructions, potentially preserving information-rich but task-irrelevant background regions. This unilateral reliance on single-modal metrics constrains further improvements in compression efficiency and accuracy.

To overcome these limitations, we propose **IFD-Prune**, a training-free, plug-and-play visual token prune framework. Unlike previous methods relying on heuristic thresholding, we reformulate the token pruning problem as a **Volumetric Information Maximization** problem (Kulesza and Taskar, 2012), utilizing iterative greedy orthogonal projection to select tokens that span the largest effective hypervolume in the feature space. This formulation empowers IFD-Prune to naturally balance sample quality and set diversity: when selecting a subset, it not only prioritizes the individual importance of each token via the weighted feature magnitude, but also penalizes redundancy between tokens via soft orthogonalization, thereby mathematically guaranteeing that the selected subset possesses the maximum feature spanning volume.

Specifically, to mitigate biases from solely relying on unimodal metrics, we design a dual quality assessment mechanism to coordinate **"Image-Intrinsic Global Context"** and **"Text-Guided Relevance"**. This mechanism utilizes the inherent attention maps of the vision encoder as a structural prior to prevent the loss of global image context due to overfitting text instructions; simultaneously, it introduces text relevance scores as guidance to enhance the saliency of task-relevant regions. Based on this hybrid quality metric, we employ an efficient **iterative greedy algorithm based on orthogonal projection** to solve the maximum volume problem (Pati et al., 1993). By eliminating redundant projection components in the feature space at each step, we extract the visual token subset with the highest information density at extremely low computational cost.

The main contributions of this paper are summarized as follows:

- We formally formulate visual token pruning as a maximum volumetric information problem, which jointly optimizes token *Quality* and set *Diversity*, addressing the imbalance in prior work that focuses on only one aspect.
- We propose IFD-Prune, a training-free and plug-and-play visual token pruning framework. We synergize a dual-criteria importance scorer with an iterative residual projection mechanism to select high-value, non-redundant tokens.
- Extensive experiments on multiple models and various standard benchmarks demonstrate that IFD-Prune achieves superior efficiency-accuracy trade-offs, outperforming recent baselines at various compression rates.

2 Related Works

2.1 Multimodal Large Language Models

In recent years, Multimodal Large Language Models (MLLMs) (Liu et al., 2024; Yang et al., 2024a; Bai et al., 2023) have achieved remarkable progress in tasks such as visual question answering and visual reasoning by aligning visual encoders with Large Language Models (LLMs) (Vaswani et al., 2017; Wei et al., 2022). Mainstream architectures, such as LLaVA and Qwen-VL (Bai et al., 2025),

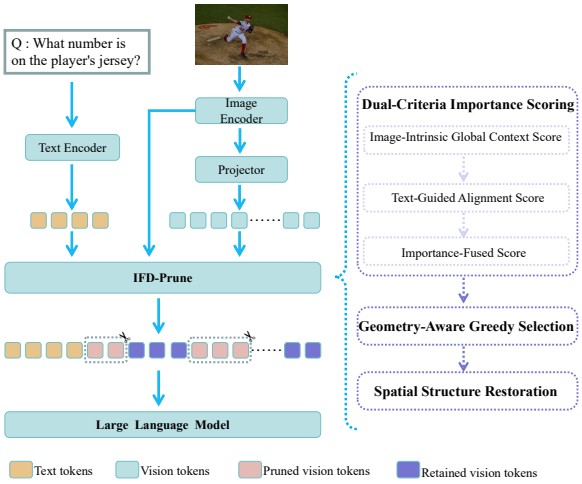


Figure 2: An overview of the MLMMs architecture, with IFD-Prune applied to visual tokens.The blocks on the right-hand side illustrate the steps of the method.

typically employ Vision Transformers (ViTs) (Radford et al., 2021; Zhai et al., 2023) to extract semantic features, which are then projected into the LLM's semantic space via linear projectors, enabling joint cross-modal reasoning. However, to enhance the model's perception of fine-grained visual details, the field is witnessing a shift towards high-resolution processing with adaptive spatial partitioning. This trend has led to a dramatic expansion of input visual sequences. While early models like LLaVA-v1.5 utilized a fixed resolution that has 576 visual tokens, recent models such as LLaVA-NeXT, Qwen2.5-VL, and Mini-Gemini-HD (Li et al., 2024b) adopt dynamic slicing or high-resolution encoding strategies. Consequently, the number of visual tokens for a single image has surged to 2,880 or even over 4,000. Given the $O(N^2)$ computational complexity of the Transformer self-attention mechanism, this explosive growth in visual sequences creates significant bottlenecks in inference latency and memory consumption. Therefore, achieving efficient visual token compression while preserving semantic fidelity has become a critical challenge that urgently needs to be addressed.

2.2 Efficient MLLMs

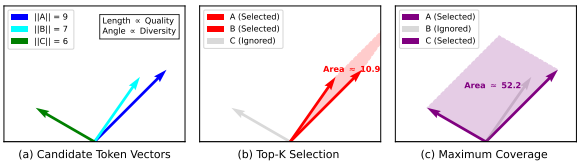
A variety of techniques have been proposed to improve the inference efficiency of Multimodal Large Language Models. These include knowledge distillation, where a compact student MLLMs is trained to mimic the behavior of a larger teacher model (Li et al., 2024a); Structured pruning, which removes redundant weights or attention heads based on importance criteria (He and Xiao, 2023); Quantization, which lowers numerical precision to reduce memory footprint and accelerate computation (Lin et al., 2023); and KV cache compression, which reduces redundant computation during autoregressive decoding by caching key-value pairs from previously generated tokens (Yuan et al., 2024).

2.3 Visual Token Pruning in MLLMs

To alleviate the computational bottleneck imposed by the massive number of visual tokens in MLLMs, existing visual token pruning methods are primarily categorized into two streams based on their utilization of textual guidance: Vision-Centric Pruning and Text-Interaction Pruning.

Vision-Centric Pruning. Methods in this category operate prior to the LLM input, relying solely on the intrinsic properties of visual features to reduce sequence length. For instance, DivPrune (Alvar et al., 2025) calculates the cosine distance between visual tokens to select a subset that maximizes semantic diversity. Similarly, VisionZip (Yang et al., 2024b) utilizes the attention weights of the CLS token in the Vision Encoder to rank patch importance, preserving high-weight tokens while fusing low-weight ones. However, the primary limitation of these approaches lies in their *task-agnostic* nature. Due to the lack of guidance from user textual instructions, the generated pruning masks remain static and fail to adapt to dynamic user queries, applying an identical strategy regardless of the input question.

Text-Interaction Pruning. The second category leverages text-visual cross-attention within the MLLMs to guide visual token pruning. For instance, SparseVLM (Zhang et al., 2024b) identifies and preserves visual tokens highly relevant to the text based on attention maps from the initial MLLMs layers, while clustering or fusing the remainder. LightVLM (Hu et al., 2025) integrates hierarchical token merging with KV Cache compression, progressively reducing visual tokens during encoding and dynamically evicting low-contribution tokens during decoding. Despite incorporating textual guidance, these methods still face significant challenges: (1) *Front-loaded computational overhead*, as they require executing partial MLLMs layers to obtain guidance signals—contradicting the goal of acceleration; and (2) *Semantic redundancy*, as greedy selection based on relevance scores often retains multiple highly similar tokens, failing to ensure representational



(a) Candidate Token Vectors (b) Top-K Selection (c) Maximum Coverage

Figure 3: Geometric comparison of token selection strategies. (a) Visual tokens are represented as vectors where length indicates quality and angle indicates diversity. (b) Traditional Top-K selection results in feature space collapse by choosing redundant vectors. (c) Volumetric Maximization: maximizes the spanned volume, effectively balancing quality and diversity for optimal information coverage.

diversity and potentially overlooking globally informative regions.

3 Method

In this section, we present IFD-Prune for efficient MLLMs inference. We first introduce the problem formulation, followed by a detailed description of the proposed framework. The overall pipeline is illustrated in Figure 2.

3.1 Problem Formulation

Multimodal Large Language Models typically encode an input image I into a sequence of visual tokens $V = [v_1, \dots, v_N] \in \mathbb{R}^{N \times d}$, where each v_i represents a local patch in a high-dimensional semantic space. These visual tokens are concatenated with textual tokens and jointly processed by the language decoder via self-attention.

However, the full set V often contains substantial semantic redundancy, and individual tokens exhibit heterogeneous utility for the downstream task. In this work, we formulate visual token pruning as a **Volumetric Information Maximization** problem. Our goal is to select a compact subset $S \subseteq \{1, \dots, N\}$ of size $K \ll N$ that maximizes the volume of the parallelotope spanned by the selected vectors in the feature space:

$$\max_{S, |S|=K} \det(\mathbf{V}_S \mathbf{V}_S^\top), \tag{1}$$

where $\mathbf{V}_S \in \mathbb{R}^{d \times K}$. Crucially, we interpret this geometric volume as a proxy for semantic information coverage. A collapsed volume implies either high linear dependence or vanishing vector magnitude, whereas a maximized volume indicates that the selected tokens are both informative and complementary, thereby spanning the widest possible semantic scope of the original image. The visual analysis is shown in the Figure 3, simplified 2D version.

- **High Semantic Utility**: Each selected token should carry significant information content. Geometrically, this corresponds to a large vector norm. We define this utility by fusing intrinsic visual saliency and cross-modal relevance.
- **High Semantic Diversity**: The selected tokens should be geometrically orthogonal to minimize information overlap. This ensures that each added token contributes a novel semantic component perpendicular to the subspace spanned by previously selected tokens.

Since finding the exact optimal subset S is NP-hard, we propose IFD-Prune to approximate this objective efficiently via a greedy strategy that iteratively selects high-utility tokens while suppressing redundancy through **residual orthogonal projection** (Pati et al., 1993; Chen et al., 2017).

3.2 Dual-Criteria Importance Scoring

To construct a quality metric that reflects both visual structure and user intent, we propose a hybrid scoring mechanism that fuses intrinsic and extrinsic cues.

Image-Intrinsic Global Context ($\mathcal{A}$): To identify image regions that contribute most significantly to the global visual representation, we extract intrinsic importance cues from the self-attention mechanism. Specifically, we utilize the attention maps from the penultimate layer (the $(L-1)$-th layer). We explicitly bypass the final layer because it serves as the direct interface to the vision-language projector. Features in the final layer often undergo a semantic shift to maximize text compatibility, potentially sacrificing the fine-grained spatial structural details required for accurate visual pruning.

The visual saliency score for the i-th token is calculated by accumulating the attention weights assigned to it by all other visual tokens. Let $\mathbf{M} \in \mathbb{R}^{N \times N}$ denote the attention matrix aggregated across all attention heads, where N is the number of visual patches. The score is defined as:

$$\mathcal{A}_i = \sum_{j=1}^{N} \mathbf{M}_{j,i}. \tag{2}$$

Intuitively, this metric quantifies the graph centrality or information density of a patch, serving as a

Figure 4: Top-64 patches selected via visual-text cosine similarity are preserved, while the remainder are masked.

robust proxy for the token's role as a semantic hub within the global visual context.

Text-Guided Alignment ($\mathcal{R}$): While intrinsic visual saliency captures the global semantic structure of the image, it is agnostic to user intent. To align token selection with the current user intent, we introduce a text-guided relevance score. Specifically, let $\mathcal{T} = \{t_1, \dots, t_M\}$ denote the set of textual token embeddings derived from the input prompt, where each $t_k \in \mathbb{R}^d$ resides in the same embedding space as the visual tokens. We define the relevance of visual token v_i as its maximum cosine similarity to any textual token:

$$\mathcal{R}_i = \max_{k \in \{1,\dots,M\}} \left(\frac{v_i \cdot t_k}{\|v_i\|_2 \, \|t_k\|_2} \right). \tag{3}$$

This follows from the observed keyword-region alignment: as shown in Figure 4, visual patches corresponding to a textual keyword yield significantly higher cosine similarities. We employ the **max** operator rather than averaging because visual regions typically exhibit *sparse alignment* with textual instructions—each object region correlates strongly with only a few relevant keywords, while showing weak similarity to the majority of the sentence.

Dual-Perspective Quality ($\mathcal{H}$): To ensure the selected tokens are both structurally informative and semantically relevant, we fuse the above metrics. First, $\mathcal{A}$ and $\mathcal{R}$ are normalized to $[0, 1]$ via Min-Max Normalization. The final quality score I_i is computed as:

$$H_i = \alpha \cdot \text{Norm}(\mathcal{A}_i) + (1 - \alpha) \cdot \text{Norm}(\mathcal{R}_i) \tag{4}$$

where α is a hyperparameter governing the trade-off. This fusion strategy achieves a dual-perspective quality assessment, simultaneously quantifying the intrinsic information density and the extrinsic alignment with user intent.

3.3 Geometry-Aware Greedy Selection

Given the hybrid importance scores from Section 3.2, a naive strategy would select the top-K visual tokens. However, as illustrated in Figure 3 (b), this approach is geometrically flawed. Magnitude alone does not guarantee coverage: even if the selected vectors possess substantial lengths, a narrow angular separation between them results in a severely diminished spanned volume. This leads to a redundant subset that fails to represent the diversity of the original semantic space.

To address this, we propose a **geometry-aware greedy selection** algorithm that iteratively constructs the subset $\mathbf{V}_S$ by balancing **semantic utility** and **directional novelty**. The procedure maintains a residual feature matrix that progressively suppresses components already captured by selected tokens.

We first initialize the residual vectors by scaling the normalized visual features with their fused importance:

$$\mathbf{Q}_i^{(0)} = \mathbf{H}_i \cdot \frac{v_i}{\|v_i\|_2} \tag{5}$$

where $\mathbf{Q}^{(0)} \in \mathbb{R}^{N \times d}$, and $\mathbf{Q}_i^{(0)}$ represents the importance-weighted feature of visual token i. We define the selected subset index as $\mathcal{I}_{select}$, starting with an empty set $\mathcal{I}_{select}^{(0)} = \varnothing$, after updating the selected vector index i_t will be added to $\mathcal{I}_{select}$.

$$i_t = \arg\max_{i \notin \mathcal{I}_{select}^{(t-1)}} \|\mathbf{Q}_i^{(t-1)}\|_2^2 \tag{6}$$

$$\mathcal{I}_{select}^{(t)} = \mathcal{I}_{select}^{(t-1)} \cup \{i_t\} \tag{7}$$

The inference proceeds in K iterations. At step t, we select the token that maximizes the residual norm index i_t, which is inspired by the maximization of the marginal volumetric gain in a Determinantal Point Process(Kulesza and Taskar, 2012; Chen et al., 2017).

After selection, we update the remaining tokens. Standard geometry-aware greedy enforces strict orthogonality ($\gamma = 1.0$), which effectively removes all features correlated with the selected token. However, visual objects often exhibit continuous semantic distribution. Strict removal can lead to object fragmentation. To address this, we introduce a **Soft Geometric Inhibition** mechanism with a relaxation factor γ:

$$\mathbf{Q}^{(t)} = \mathbf{Q}^{(t-1)} - \gamma \cdot \frac{\mathbf{Q}^{(t-1)} \cdot (\mathbf{Q}_{i_t}^{(t-1)})^\top}{\|\mathbf{Q}_{i_t}^{(t-1)}\|_2^2} \mathbf{Q}_{i_t}^{(t-1)} \tag{8}$$

where $\gamma \in (0, 1]$ controls the strength of redundancy suppression. We empirically set $\gamma =$

Method	Venue	GQA	MMB	MME	POPE	SQA^{Img}	VQA^{v2}	VQA^{Text}	SEED	VizWiz	Avg
Vanilla, 576 Tokens (100%)											
LLaVA-1.5-7B	CVPR'24	61.90	64.70	1862.0	85.90	69.50	78.50	58.20	58.60	50.80	100.0%
Retain 192 Tokens (33.3%)											
ToMe	ICLR'23	54.30	60.50	1563.0	72.40	65.20	68.00	52.10	–	–	88.49%
FastV	ECCV'24	52.70	61.20	1612.0	64.80	67.30	67.10	52.50	57.10	50.80	90.19%
HiRED	AAAI'25	58.80	62.60	1742.0	83.00	67.90	75.00	–	–	51.10	96.54%
TRIM	ACL'25	59.90	64.10	1765.0	87.10	67.80	76.20	54.90	–	50.40	97.52%
PyramidDrop	CVPR'25	57.30	63.30	1797.0	82.30	69.00	75.10	56.50	–	51.10	96.92%
VisionZip	CVPR'25	59.30	63.00	1783.0	85.30	68.90	76.80	57.30	56.40	–	97.49%
SparseVLM	ICML'25	59.50	64.10	1780.0	85.40	68.80	77.00	57.70	55.80	50.60	97.92%
DivPrune	CVPR'25	59.97	62.54	1762.2	87.00	68.91	76.87	56.97	58.71	–	98.08%
HiPrune	AAAI'26	59.20	62.80	1814.0	86.10	68.90	76.70	57.60	–	54.50	99.18%
IFD-Prune	–	59.86	63.92	1803.2	87.91	68.62	75.11	57.48	59.19	54.49	**99.57%**
Retain 128 Tokens (22.2%)											
ToMe	ICLR'23	52.40	53.30	1343.0	62.80	59.60	63.00	49.10	–	50.20	82.68%
FastV	ECCV'24	49.60	56.10	1490.0	59.60	60.20	61.80	50.60	55.90	51.30	84.99%
HiRED	AAAI'25	57.10	61.70	1714.0	79.80	68.10	73.50	–	–	51.40	95.05%
TRIM	ACL'25	58.90	63.30	1732.0	87.20	68.40	74.80	52.70	–	50.60	96.42%
PyramidDrop	CVPR'25	57.10	61.60	1761.0	82.30	68.40	72.90	56.60	–	51.00	95.85%
VisionZip	CVPR'25	57.60	62.00	1762.0	83.20	68.90	75.60	56.80	54.90	–	95.89%
SparseVLM	ICML'25	53.80	64.40	1761.0	85.00	68.50	76.30	56.70	53.40	50.20	95.90%
DivPrune	CVPR'25	59.25	62.03	1718.2	86.72	68.96	75.96	56.06	56.98	–	96.80%
HiPrune	AAAI'26	57.30	62.20	1782.0	82.80	68.30	74.90	56.64	–	54.30	97.34%
IFD-Prune	–	59.26	62.89	1748.1	87.90	69.21	73.97	56.81	58.35	54.67	**98.64%**
Retain 64 Tokens (11.1%)											
ToMe	ICLR'23	48.60	43.70	1138.0	52.50	50.00	57.10	45.30	–	50.20	73.70%
FastV	ECCV'24	46.10	48.00	1256.0	48.00	51.10	55.00	47.80	51.90	50.80	76.25%
HiRED	AAAI'25	54.60	60.20	1595.0	73.70	68.20	69.80	–	–	53.30	92.10%
TRIM	ACL'25	56.90	61.50	1603.0	86.70	69.00	71.90	50.00	–	50.60	93.80%
PyramidDrop	CVPR'25	47.50	58.80	1561.0	55.90	69.20	69.20	50.60	–	50.70	86.37%
VisionZip	CVPR'25	55.10	60.10	1690.0	77.00	69.00	72.40	55.50	52.20	–	92.28%
SparseVLM	ICML'25	53.70	60.10	1559.0	77.50	69.70	70.20	53.40	51.10	50.40	91.27%
DivPrune	CVPR'25	57.78	59.28	1674.4	85.56	68.17	74.11	54.69	55.13	–	94.38%
HiPrune	AAAI'26	53.60	59.50	1646.0	73.00	68.90	69.20	54.90	–	54.40	92.58%
IFD-Prune	–	57.16	61.77	1709.2	87.14	68.17	71.72	54.84	56.92	55.85	**96.87%**

Table 1: **Performance comparison of different pruning methods on LLaVA-1.5-7B.** Avg represents the average percentage of performance maintained at the corresponding reduction ratio. All baseline are training-free. All methods data are collected from previously published methods.

0.8. This generalized update rule suppresses background redundancy while preserving the structural integrity of salient objects. The process repeats until $|\mathcal{I}_{select}| = K$.

3.4 Spatial Structure Restoration

A critical oversight in prior visual pruning methods is the neglect of sequential integrity. Simply feeding top-K tokens in descending order of importance disrupts the raster-scan order learned by the ViT's positional embeddings, leading to **Positional Misalignment** in the MLLMs.

To remedy this, we decouple the selection criterion from the input order. Let $\mathcal{I}_{select}$ be the set of indices selected by our algorithm. We perform an index sorting operation:

$$\mathcal{I}_{sorted} = \mathrm{Sort}(\mathcal{I}_{select}, \text{ascending}) \tag{9}$$

$$\mathbf{V}_S = \{v_i \mid i \in \mathcal{I}_{sorted}\} \tag{10}$$

The final sequence is reconstructed as $\mathbf{V}_S$. This simple step restores the underlying 2D manifold structure, ensuring the preservation of spatial causality for accurate reasoning.

4 Experiments

To comprehensively validate the effectiveness and generalizability of our proposed IFD-Prune method, we conduct extensive experiments across multiple representative MLLMs architectures and diverse benchmarks. We compare our method against various state-of-the-art visual token pruning techniques to demonstrate its superiority in balancing efficiency and performance.

4.1 Experimental Settings

Models. As a Training-free and Plug-and-play method, IFD-Prune is designed to be compatible with generic MLLMs architectures comprising a vision encoder and a Large Language Model. To

verify this universality, we select four models covering different parameter scales and visual encoding strategies: including the static-resolution LLaVA-v1.5 (7B and 13B), LLaVA-NeXT-7B, and the dynamic-resolution Qwen2.5-VL-7B-Instruct.

Benchmarks. We categorize the datasets into three core capability domains for performance evaluation:(1) Visual Question Answering: including VQAv2 (Goyal et al., 2016), VizWiz (Gurari et al., 2018), GQA (Hudson and Manning, 2019), TextVQA (Singh et al., 2019), and ScienceQA (Lu et al., 2022).(2) Comprehensive Competence: including MME (Fu et al., 2023), MMBench (Liu et al., 2023b), and SEED-Bench (Li et al., 2023b).(3) Hallucination Detection: utilizing the POPE (Li et al., 2023c) benchmark to verify whether token pruning exacerbates object hallucination issues.

Baselines. We compare against 10 mainstream SOTA visual token pruning methods, including: ToMe (Bolya et al., 2022), FastV (Chen et al., 2024), HiRED (Arif et al., 2024), TRIM (Song et al., 2025), PyramidDrop (Xing et al., 2024), VisionZip (Yang et al., 2024b), SparseVLM (Zhang et al., 2024b), DivPrune (Alvar et al., 2025), MMTok (Dong et al., 2025) and HiPrune (Liu et al., 2025). Notably, limited by code availability or architectural compatibility, certain baselines cannot be directly applied to LLaVA-NeXT or Qwen2.5-VL. In such cases, we compare IFD-Prun: with the subset of baselines compatible with that specific architecture to ensure the fairness and rigor of the evaluation.

Implementation Details. For all baseline methods, we strictly follow the default configurations provided in their official codebases. For IFD-Prune, we uniformly set the hyperparameters to $\alpha = 0.8$ and $\gamma = 0.9$ across all models. All evaluations are conducted based on the lmms-eval framework (Zhang et al., 2024a), and FLOPs metrics are calculated using the calflops toolkit. All experiments are performed on a single NVIDIA H20 GPU.

4.2 Main Results

Results in LLaVA. We first evaluate IFD-Prune on the LLaVA-v1.5-7B model, with detailed comparative results presented in Table 1. Following mainstream settings, we evaluate performance under budgets retaining 192 (33.3%), 128 (22.2%), and 64 (11.1%) visual tokens, respectively. Results demonstrate that IFD-Prune consistently exhibits state-of-the-art performance across all pruning ratios. Notably, IFD-Prune achieves near-lossless visual token pruning: with **66.7%** of tokens removed, it preserves **99.57%** of the original baseline performance; even when **77.8%** of tokens are discarded, it maintains **98.64%** of the performance. Under the most aggressive **88.9%** pruning rate (retaining only 64 tokens), it still retains **96.87%** of the performance. This result fully validates the effectiveness of the IFD-Prune strategy in capturing high-value semantic information under extremely low token budgets.

Method	GQA	MMB	MME	POPE	VQA^{Text}	Avg(%)
Vanilla, 576 Tokens (100%)						
LLaVA	63.20	67.70	1818.00	85.90	61.30	100.0%
Retain 192 Tokens (33.3%)						
VisionZip	59.10	66.90	1754.00	85.10	59.50	96.99%
DivPrune	59.42	66.58	1781.50	86.76	58.46	97.35%
HiPrune	59.40	67.10	1798.00	85.40	59.50	97.70%
IFD-Prune	59.81	66.67	1772.93	88.09	59.40	**98.02%**
Retain 128 Tokens (22.2%)						
VisionZip	57.90	66.70	1743.00	85.20	58.70	96.19%
DivPrune	58.89	66.07	1748.56	86.53	58.17	96.52%
HiPrune	57.90	66.80	1730.00	82.80	58.30	95.39%
IFD-Prune	59.00	66.07	1760.61	88.02	58.41	**97.11%**
Retain 64 Tokens (11.1%)						
VisionZip	58.10	65.60	1671.00	81.60	58.50	94.23%
DivPrune	57.66	64.60	1777.93	84.80	57.11	**95.27%**
HiPrune	54.20	64.80	1634.00	72.40	56.70	89.63%
IFD-Prune	57.34	64.00	1699.59	87.14	57.69	94.86%

Table 2: **Results on LLaVA-v1.5-13B.**

Method	GQA	MMB	MME	POPE	VQA^{Text}	Avg(%)
Vanilla, 2880 (576*5) Tokens (100%)						
LLaVA	64.20	67.90	1842.00	86.40	61.30	100.00%
Retain 640 Tokens (22.2%)						
VisionZip	61.30	66.30	1787.00	86.30	60.20	97.65%
SparseVLM	60.30	65.70	1772.00	85.30	57.80	95.98%
DivPrune	61.58	65.38	1773.04	85.51	55.41	95.57%
IFD-Prune	62.17	65.89	1801.76	88.66	58.88	**98.07%**
Retain 320 Tokens (11.1%)						
VisionZip	59.30	63.10	1702.00	82.10	58.90	93.76%
SparseVLM	57.70	64.30	1694.00	76.90	55.90	91.35%
DivPrune	59.63	63.66	1731.04	83.47	53.84	93.01%
IFD-Prune	60.65	63.75	1793.44	88.09	57.63	**96.34%**
Retain 160 Tokens (5.6%)						
VisionZip	55.50	60.10	1630.00	74.80	58.30	89.03%
SparseVLM	51.20	63.10	1542.00	-	46.4	85.46%
DivPrune	57.79	62.29	1658.25	79.36	52.42	89.83%
IFD-Prune	58.86	63.23	1698.09	87.47	54.75	**93.51%**

Table 3: **Results on LLaVA-NeXT-7B.**

Further extending our evaluation to LLaVA-v1.5-13B (Table 2) and LLaVA-NeXT-7B (Table 3), IFD-Prune demonstrates exceptional scalability and adaptability. On the LLaVA-v1.5-13B model, performance remains robust, retaining **94.86%** of the baseline capability even with **88.9%** of visual tokens removed. For LLaVA-NeXT-7B, given that its high-resolution strategy can generate up to 2,880

Method	GQA	MMB	MME	POPE	VQA^{Text}	Avg(%)
Vanilla, 100% Tokens						
Qwen2.5	60.50	83.20	2331.00	86.20	83.10	100.00%
Retain 20% Tokens						
VisionZip	56.80	80.33	2174.00	83.38	70.43	93.04%
DivPrune	56.70	76.98	2163.00	80.59	65.86	90.36%
MMTok	58.09	79.30	2217.00	82.38	70.49	93.37%
IFD-Prune	57.49	80.84	2260.86	85.88	77.95	**96.52%**
Retain 10% Tokens						
VisionZip	52.47	75.60	2003.00	78.90	63.78	86.36%
DivPrune	53.43	72.85	1957.00	74.99	59.59	83.71%
MMTok	55.09	74.74	2051.00	78.75	63.90	87.43%
IFD-Prune	53.78	76.29	2107.01	83.31	70.95	**90.60%**
Retain 5% Tokens						
VisionZip	46.28	67.53	1677.00	66.38	54.49	74.44%
DivPrune	49.01	65.89	1739.00	68.45	52.02	75.36%
MMTok	50.66	65.89	1796.00	71.35	55.95	78.38%
IFD-Prune	48.67	70.70	1814.83	77.86	58.30	**80.75%**

Table 4: **Results on Qwen2.5-VL-7B-Instruct.**

Method	GQA	MMB	MME	POPE	VQA^{T}	Avg(%)
w/o T-R	56.03	60.40	1702.42	83.34	56.01	98.64%
w/o V-S	56.70	60.22	1666.11	87.09	52.94	98.13%
w/o Div	53.81	59.36	1638.55	84.28	52.19	95.60%
w/o Sort	57.22	61.17	1702.53	86.72	54.13	99.39%
Full Method	57.16	61.77	1709.20	87.14	54.84	**100.00%**

Table 5: **Ablation study.** All settings are evaluated on the LLaVA-v1.5-7B model and retain 64 visual tokens. "w/o T-R" denotes setting $\alpha = 0$, "w/o V-S" denotes setting $\alpha = 1$, "w/o Div" denotes replacing our orthogonalization strategy with top-k selection, "w/o Sort" denotes without spatial structure restoration.

visual tokens, we achieve even more aggressive visual token pruning, in the extreme case of retaining only **5.6%** (160 tokens), the model still maintains **93.51%** of its average performance. This highlights the capability of IFD-Prune to effectively filter out the massive redundancy within visual tokens.

Results in Qwen. To verify the adaptability of our method to dynamic resolution architectures, we extend our evaluation to Qwen2.5-VL-7B, with results summarized in Table 4. In terms of average performance, IFD-Prune consistently outperforms all competing baselines across various pruning budgets. Specifically, with a **20%** visual token budget, our method retains **96.52%** of the original model's average performance, surpassing the strongest baseline by over **3.1%**. As the pruning rate increases, the performance gap becomes even more pronounced. Under the rigorous budget of retaining only **10%** of visual tokens, IFD-Prune preserves **90.60%** of the baseline capability, establishing a clear lead over VisionZip (86.36%) and DivPrune (83.71%). Even in the extreme scenario of retaining merely **5%** of tokens, our method remains robust, retaining **80.75%** of the original performance, which validates its effectiveness in handling MLLMs with dynamic resolution architectures.

4.3 Ablation Studies

We conduct ablation studies on the individual components of IFD-Prune, with results summarized in Table 5. Overall, the removal of any component leads to a degradation in average performance, verifying the necessity of the proposed modules.When the diversity constraint is removed, relying solely on a Top-K selection strategy, we observe the most significant performance drop, with a decrease of **4.4%** in average performance. This validates that employing only the Top-K strategy leads to information redundancy, thereby impairing performance.Removing the query-guided relevance score results in a **1.36%** drop in average performance. Notably, the score on the POPE benchmark declines significantly in this setting, demonstrating that query-guided relevance is effective in mitigating hallucinations. Removing the Image-Intrinsic Global Context leads to a **1.87%** decline in performance.Finally, removing the re-ranking strategy results in a **0.61%** drop, validating that maintaining the original positional relationships among visual tokens aids the LLM in better comprehending the visual information.

5 Conclusion

In this paper, we propose IFD-Prune, a token pruning method for vision-language models based on maximizing volumetric information. IFD-Prune balances selection quality and diversity. Quality is driven by a dual-criteria mechanism fusing visual saliency and cross-modal textual relevance, while diversity is guaranteed via iterative greedy orthogonal projection to select tokens spanning the largest hypervolume in feature space. Without fine-tuning, IFD-Prune maintains strong performance even at high pruning ratios. Extensive experiments on multiple models and benchmarks show that IFD-Prune achieves state-of-the-art accuracy, generalizes well across model sizes/architectures, and reduces memory consumption and end-to-end latency for MLMMs.

Limitations

Currently, our method applies a uniform pruning ratio across all inputs. However, the information density varies significantly between complex and simple samples. An instance-aware adaptive pruning strategy, which dynamically determines the optimal token budget for each image, could further enhance the efficiency-performance trade-off.

Our approach relies on hyperparameters that are empirically optimized. While robust across tested datasets, these parameters are fixed. We recognize that a lightweight few-shot calibration or data-driven tuning phase on domain-specific data could potentially yield superior performance compared to the current training-free configuration.

Our validation is currently confined to static Image-Text tasks. Given the significant information redundancy inherent in video streams, extending IFD-Prune to Video-LLMs to tackle spatiotemporal redundancy remains a promising avenue for future exploration.

References

OpenAI Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haim ing Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, and 260 others. 2023. Gpt-4 technical report.

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 Jin. 2024. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models in resource-constrained environments. *ArXiv*, abs/2408.10945.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. 2023. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond.

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2022. Token merging: Your vit but faster. *ArXiv*, abs/2210.09461.

Laming Chen, Guoxin Zhang, and Hanning Zhou. 2017. Improving the diversity of top-n recommendation via determinantal point process. *ArXiv*, abs/1709.05135.

Liang Chen, Haozhe Zhao, Tianyu Liu, Shuai Bai, Junyang Lin, Chang Zhou, and Baobao Chang. 2024. 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*.

Sixun Dong, Juhua Hu, Mian Zhang, Ming Yin, Yanjie Fu, and Qi Qian. 2025. Mmtok: Multimodal coverage maximization for efficient inference of vlms. *ArXiv*, abs/2508.18264.

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.

Yash Goyal, Tejas Khot, Douglas Summers-Stay, Dhruv Batra, and Devi Parikh. 2016. Making the v in vqa matter: Elevating the role of image understanding in visual question answering. *International Journal of Computer Vision*, 127:398 – 414.

Danna Gurari, Qing Li, Abigale Stangl, Anhong Guo, Chi Lin, Kristen Grauman, Jiebo Luo, and Jeffrey P. Bigham. 2018. Vizwiz grand challenge: Answering visual questions from blind people. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3608–3617.

Yang He and Lingao Xiao. 2023. Structured pruning for deep convolutional neural networks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46:2900–2919.

Lianyu Hu, Fanhua Shang, Wei Feng, and Liang Wan. 2025. Lightvlm: Acceleraing large multimodal models with pyramid token merging and kv cache compression. *ArXiv*, abs/2509.00419.

Drew A. Hudson and Christopher D. Manning. 2019. Gqa: a new dataset for compositional question answering over real-world images. *ArXiv*, abs/1902.09506.

Alex Kulesza and Ben Taskar. 2012. Determinantal point processes for machine learning. *Found. Trends Mach. Learn.*, 5:123–286.

Bohao Li, Yuying Ge, Yixiao Ge, Guangzhi Wang, Rui Wang, Ruimao Zhang, and Ying Shan. 2023a. Seed-bench-2: Benchmarking multimodal large language models. *ArXiv*, abs/2311.17092.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023b. Seed-bench: Benchmarking multimodal llms with generative comprehension. *ArXiv*, abs/2307.16125.

Xiang Li, Shizhu He, Jiayu Wu, Zhao Yang, Yao Xu, Yang jun Jun, Haifeng Liu, Kang Liu, and Jun Zhao. 2024a. MoDE-CoTD: Chain-of-thought distillation for complex reasoning tasks with mixture of decoupled LoRA-experts. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 11475–11485, Torino, Italia. ELRA and ICCL.

Yanwei Li, Yuechen Zhang, Chengyao Wang, Zhisheng Zhong, Yixin Chen, Ruihang Chu, Shaoteng Liu, and Jiaya Jia. 2024b. Mini-gemini: Mining the potential of multi-modality vision language models. *IEEE transactions on pattern analysis and machine intelligence*, PP.

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

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2024. Video-LLaVA: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5971–5984, Miami, Florida, USA. Association for Computational Linguistics.

Ji Lin, Jiaming Tang, Haotian Tang, Shang Yang, Xingyu Dang, and Song Han. 2023. Awq: Activation-aware weight quantization for on-device llm compression and acceleration. *GetMobile: Mobile Computing and Communications*, 28:12 – 17.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 26286–26296.

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

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*, abs/2508.00553.

Yuanzhan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2023b. Mmbench: Is your multi-modal model an all-around player? In *European Conference on Computer Vision*.

Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and A. Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *ArXiv*, abs/2209.09513.

Y.C. Pati, R. Rezaiifar, and P.S. Krishnaprasad. 1993. Orthogonal matching pursuit: recursive function approximation with applications to wavelet decomposition. In *Proceedings of 27th Asilomar Conference on Signals, Systems and Computers*, pages 40–44 vol.1.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*.

Amanpreet Singh, Vivek Natarajan, Meet Shah, Yu Jiang, Xinlei Chen, Dhruv Batra, Devi Parikh, and Marcus Rohrbach. 2019. Towards vqa models that can read. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8309–8318.

Dingjie Song, Wenjun Wang, Shunian Chen, Xidong Wang, Michael X. Guan, and Benyou Wang. 2025. Less is more: A simple yet effective token reduction method for efficient multi-modal LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7614–7623, Abu Dhabi, UAE. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aur'elien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *ArXiv*, abs/2302.13971.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Neural Information Processing Systems*.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *ArXiv*, abs/2206.07682.

Long Xing, Qidong Huang, Xiao wen Dong, Jiajie Lu, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, Jiaqi Wang, Feng Wu, and Dahua Lin. 2024. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *ArXiv*, abs/2410.17247.

An Yang, Baosong Yang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Zhou, Chengpeng Li, Chengyuan Li, Dayiheng Liu, Fei Huang, Guanting Dong, Haoran Wei, Huan Lin, Jialong Tang, Jialin Wang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Ma, and

39 others. 2024a. Qwen2 technical report. *ArXiv*, abs/2407.10671.

Senqiao Yang, Yukang Chen, Zhuotao Tian, Chengyao Wang, Jingyao Li, Bei Yu, and Jiaya Jia. 2024b. Visionzip: Longer is better but not necessary in vision language models. *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 19792–19802.

Jiayi Yuan, Hongyi Liu, Shaochen Zhong, Yu-Neng Chuang, Songchen Li, Guanchu Wang, Duy Le, Hongye Jin, Vipin Chaudhary, Zhaozhuo Xu, Zirui Liu, and Xia Hu. 2024. Kv cache compression, but what must we give in return? a comprehensive benchmark of long context capable approaches. In *Conference on Empirical Methods in Natural Language Processing*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11941–11952.

Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Yuhao Dong, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2024a. Lmms-eval: Reality check on the evaluation of large multimodal models. In *North American Chapter of the Association for Computational Linguistics*.

Yuan Zhang, Chunkai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and Shanghang Zhang. 2024b. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *ArXiv*, abs/2410.04417.

A Efficiency Analysis

We evaluate the computational impact of IFD-PRUNE by comparing its overhead against the savings in the Transformer backbone.

A.1 Transformer Backbone Complexity

The computational cost of a standard MLLM is dominated by T Transformer layers. For each layer, the floating-point operations (FLOPs) are approximated as:

$$\text{FLOPs}_{\text{trans}} \approx T \times \left(4nd^2 + 2n^2d + 2ndm\right)$$

where:

- n is the total length ($n = n_{\text{sys}} + n_{\text{img}} + n_{\text{que}}$);
- d is the hidden dimension;
- m is the intermediate size of the FFN;

A.2 Pruning Algorithm Overhead

The computational overhead of IFD-PRUNE involves cross-modal alignment and iterative residual selection, K is the number of retained visual tokens. The total FLOPs are expressed as:

$$\text{FLOPs}_{\text{prune}} \approx 2n_{\text{img}}n_{\text{que}}d + 7Kn_{\text{img}}d + 4n_{\text{img}}d$$

A.3 Theoretical Efficiency Gain

To illustrate the efficiency benefits, we analyze **LLaVA-v1.5-7B** under an 88.9% pruning ratio, where n_{img} is reduced from 576 to 64. The model comprises a CLIP-ViT-L/14 vision encoder ($T = 14$ layers, $d = 1024$, $m = 4096$) and a Llama-2-7B backbone ($T = 32$ layers, $d = 4096$, $m = 11008$).

Assuming a typical scenario where $n_{\text{sys}} + n_{\text{que}} \approx 50$, the total input sequence length n to the LLM decreases from 626 to 114. While the computational cost of the CLIP-ViT-L/14 vision encoder remains constant, the FLOPs of the Llama-2-7B language model are drastically reduced.The total inference computation drops from 3.444 TFLOPs to 0.768 TFLOPs, representing an overall reduction of approximately 77.7% in theoretical FLOPs.The additional computation introduced by IFD-Prune is only 1.3 GFLOPs (≈ 0.0013 TFLOPs). This overhead is negligible, accounting for less than 0.05% of the total saved computation.

This analysis confirms that by effectively mitigating the n_{img} bottleneck, IFD-PRUNE achieves a superior efficiency-accuracy trade-off. It significantly accelerates the dominant language modeling phase with minimal pre-processing cost, all while maintaining 96.87% of the original model's performance.

B Extended Experiment Results

To further validate the effectiveness of the proposed IFD-PRUNE, we evaluate it on four challenging datasets—SQA, VQAv2, SEED-Bench, and VizWiz—to cover a broad range of visual reasoning capabilities. Results on LLaVA-v1.5-13B and LLaVA-v1.6-7B under varying pruning ratios are summarized in Table 6.

C Pseudo-code for IFD-Prune

As shown in Algorithm 1, we provide a pseudo-code for IFD-Prune.

method	visual token	GQA	MMB	MME	POPE	SQA^{Img}	VQA^{v2}	VQA^{Text}	SEED	VizWiz	Avg
LLaVA-v1.5-13B	576	63.20	67.70	1818.00	85.90	72.80	80.00	61.30	66.90	53.60	100.0%
IFD-Prune	192	59.81	66.67	1772.93	88.09	73.13	76.23	59.40	65.78	55.09	98.55%
	128	59.00	66.07	1760.61	88.02	72.78	75.24	58.41	64.85	55.58	97.80%
	64	57.34	64.00	1699.59	87.14	71.64	73.21	57.69	62.82	56.51	95.95%
LLaVA-NeXT-7B	2880	64.20	67.90	1842.00	86.40	70.20	80.10	61.30	70.20	55.20	100.0%
IFD-Prune	640	62.17	65.89	1801.76	88.66	68.42	77.18	58.88	68.47	58.60	98.65%
	320	60.65	63.75	1793.44	88.09	67.48	74.84	57.63	66.57	57.19	96.63%
	160	58.86	63.23	1698.09	87.47	67.72	72.09	54.75	64.69	57.24	94.43%

Table 6: **Results on LLaVA-v1.5-13B and LLaVA-NeXT-7B.**

Algorithm 1 IFD-Prune: Token Pruning Method

Require: Visual features $V \in \mathbb{R}^{N \times D}$, Text embeddings $T \in \mathbb{R}^{L \times D}$, Attention scores $A \in \mathbb{R}^{N}$, Target count K, Decay factor γ, Fusion weight α.

Ensure: Pruned visual tokens V_{sel}.

$R_i \leftarrow \max_j \text{cosine_sim}(V_i, T_j), \quad \forall i \in \{1 \dots N\}$
$H \leftarrow \alpha \cdot \text{Norm}(A) + (1 - \alpha) \cdot \text{Norm}(R)$
$\hat{V}_i \leftarrow V_i / \|V_i\|_2, \quad \forall i \in \{1 \dots N\}$
$Q_i \leftarrow \hat{V}_i \cdot H_i, \quad \forall i \in \{1 \dots N\}$
$q_i \leftarrow Q_i, \quad \forall i \in \{1 \dots N\}$
$\mathcal{I} \leftarrow \emptyset$
while $|\mathcal{I}| < K$ **do**
  $idx \leftarrow \text{argmax}_{i \notin \mathcal{I}} \|q_i\|_2^2$
  $\mathcal{I} \leftarrow \mathcal{I} \cup \{idx\}$
  $\mathbf{b} \leftarrow q_{idx}$
  if $\|\mathbf{b}\|_2^2 < \epsilon$ **then**
    break
  end if
  for all $j \notin \mathcal{I}$ **do**
    $c_j \leftarrow (q_j \cdot \mathbf{b}) / \|\mathbf{b}\|_2^2$
    $q_j \leftarrow q_j - \gamma \cdot c_j \mathbf{b}$
  end for
end while
return V at indices $\text{sort}(\mathcal{I})$

D AI Assistants Usage

We used artificial intelligence tools to assist with language polishing during the writing process and to help identify and correct errors in code. All results were reviewed and verified by the authors.