TrimTokenator: Towards Adaptive Visual Token Pruning for Large Multimodal Models

Anonymous ACL submission

Abstract

Large Multimodal Models (LMMs) have achieved significant success across various tasks. These models usually encode visual inputs into dense token sequences, which are then concatenated with textual tokens and jointly processed by a language model. However, the increased token count substantially raises computational and memory costs during inference. Token pruning has emerged as a promising approach to address this issue. Existing token pruning methods often rely on costly calibration or suboptimal importance metrics, leading to redundant retained tokens. In this paper, we analyze the redundancy differences between visual and textual tokens and propose pruning exclusively on visual tokens. Based on this, we propose a visual token pruning strategy that explicitly preserves both cross-modal alignment and intra-modal informational diversity. We introduce a mutual information-based token pruning strategy that removes visual tokens semantically misaligned with textual tokens, effectively preserving the alignment between the visual and textual modalities. We further refine the retained tokens by maximizing their expected pairwise distances in the latent space to enhance representational quality and reduce redundancy. which is solved efficiently with a greedy algorithm. Extensive experiments demonstrate that our method maintains strong performance while reducing tokens by 88.9% on models such as LLaVA-1.5-7B and LLaVA-NEXT-7B, resulting in a 56.7% improvement in inference speed.

1 Introduction

Large Multimodal Models (LMMs) (Bai et al., 2025; Team et al., 2025; Zhu et al., 2025; Liu et al., 2024a; Lin et al., 2023) have substantially enhanced the reasoning capabilities of Large Language Models (LLMs) (Brown et al., 2020; Touvron et al., 2023; Chiang et al., 2023; Zhu et al., 2023; Li et al., 2023a; Zhang et al., 2023; Huang et al., 2023; Wang et al., 2023) by enabling joint processing of multimodal inputs such as images and texts. Typically, visual inputs are encoded into dense token sequences via a vision encoder and concatenated with textual tokens for unified processing by the language model. However, the resulting token sequences often reach thousands in length, leading to significant computational and memory overhead due to the quadratic complexity of self-attention with respect to sequence length (Vaswani et al., 2017; Face, 2024; Chen et al., 2023; Keles et al., 2023; Liu et al., 2022). These limitations present major obstacles to deploying LMMs in resource-constrained or latency-sensitive environments (Chen et al., 2024a; Lin et al., 2025a).

Recent studies have shown that visual token representations in LMMs exhibit substantial redundancy (Liu et al., 2024c; Shang et al., 2024a; Huang et al., 2024; Tong et al., 2025; Li et al., 2025a). Leveraging this insight, visual token pruning methods have been proposed to reduce computational cost by selectively removing unnecessary tokens. By eliminating redundant visual tokens, these approaches effectively alleviate the quadratic burden of long input sequences. Notably, prior work demonstrates that pruning the majority of visual tokens can yield significant efficiency gains with minimal performance degradation (Zhang et al., 2024a; Chen et al., 2024a; Lin et al., 2025a; Huang et al., 2024; Sun et al.).

While token pruning is beneficial, its application to LMMs remains challenging. Existing token pruning methods can be broadly categorized into three types. 1) A common approach is to use attention scores to identify redundant tokens (Chen et al., 2024b; Lin et al., 2025b; Shang et al., 2024b; Tong et al., 2025). However, such methods are susceptible to positional bias and often retain spatially adjacent tokens with high similarity, leading to redundancy and performance degradation. 2) Other methods rely on model-specific calibration or fine-

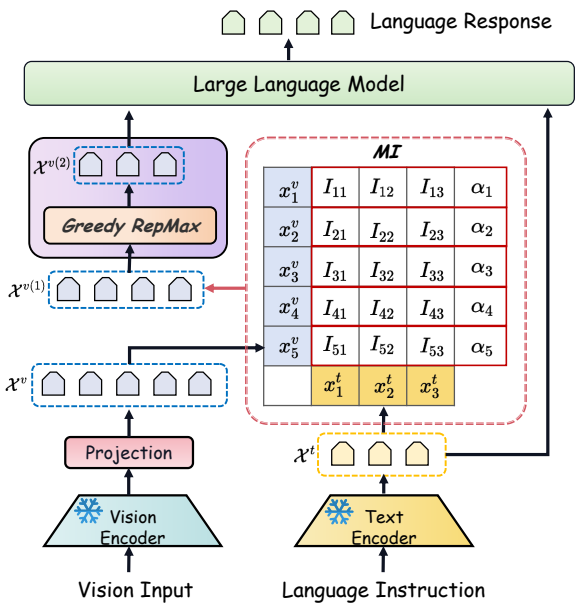

Figure 1: Overview of the visual token pruning method. We compute the average mutual information (MI) between each visual token and textual tokens to obtain its semantic alignment score α_i, and preserve the highest score tokens to form $\mathcal{X}^{v(1)}$. This subset is further refined by Greedy RepMax, which prunes redundant tokens to yield $\mathcal{X}^{v(2)}$. Greedy RepMax is a greedy approximation to the NP Hard problem of maximizing the expected pairwise distance among visual tokens.

tuning (Ye et al., 2025a; Lin et al., 2025b; Li et al., 2025c; Cai et al., 2024), which incurs high computational cost and limits scalability in practical deployment. 3) Another line of work addresses token pruning by maximizing the minimum distance between tokens (Alvar et al., 2025). While this strategy encourages token separation, it is sensitive to outliers and may fail to preserve cross-modal alignment, ultimately affecting downstream performance.

In view of these challenges and opportunities, we propose a visual token pruning strategy that explicitly preserves both cross-modal alignment and intra-modal informational diversity. Our analysis begins by examining the inherent redundancy differences between visual and textual tokens through three key perspectives: attention mechanisms, semantic distribution, and information repetition. These analyses reveal that visual tokens exhibit significantly higher redundancy compared to their textual counterparts. Consequently, we choose to prune only visual tokens while retaining all textual tokens. To preserve cross-modal alignment during pruning, we estimate the mutual information between visual and textual token embeddings. Visual tokens with higher expected mutual information with textual tokens are more likely to exhibit semantic alignment with the textual modality and are therefore retained, while those with lower mutual information are pruned as misaligned. Additionally, we retain visual tokens by maximizing their expected pairwise distances in the embedding space, thereby promoting intra-modal information diversity and encouraging rich, non-overlapping visual semantic representations. To address this problem efficiently, we employ a greedy algorithm that iteratively selects tokens exhibiting maximal dissimilarity. Extensive experiments demonstrate that our method maintains strong performance while reducing tokens by 88.9% on models such as LLaVA-1.5-7B and LLaVA-NEXT-7B, resulting in a 56.7% improvement in inference speed. In summary, our contributions can be summarized as follows:

- We analyze the redundancy differences between visual and textual tokens and propose pruning exclusively on visual tokens. Based on this, we design a visual token pruning method that explicitly preserves cross-modal alignment and intra-modal information diversity, improving inference efficiency while maintaining semantic integrity.

- We introduce a mutual information based pruning strategy that retains semantically aligned visual tokens with textual tokens while removing misaligned ones, effectively preserving the alignment between visual and textual modalities.

- We propose an information diversity driven pruning strategy that maximizes the expected pairwise distances among visual tokens in the embedding space, effectively reducing redundancy and enhancing intra modal information richness, which is solved efficiently with a greedy algorithm.

2 Related Work

2.1 Large Multimodal Models (LMMs)

Large Multimodal Models (LMMs) extend pretrained LLMs by integrating multiple modalities (Bai et al., 2025; Team et al., 2025), such as images and texts. Typically, a vision encoder extracts dense visual features, which are projected into the language model's embedding space via modules like Q-Former (Chebotar et al., 2023) or

MLPs (Taud and Mas, 2017). Conventional methods resize high-resolution images to a fixed scale (Koonce, 2021), causing geometric distortions and loss of fine-grained spatial details. Dynamic tiling mitigates this by splitting images into smaller regions, each independently encoded by a shared encoder, better preserving local information (Yuan et al., 2021; Yin et al., 2022). However, this increases the number of visual tokens and computational cost, a challenge further amplified in video LMMs due to multi-frame processing (Liu et al., 2024c). These challenges highlight the need for efficient inference methods to enable LMMs deployment under resource constraints (Tong et al., 2025).

2.2 Visual Token Pruning

Visual Token Pruning aims to reduce computational overhead and improve inference efficiency by removing redundant or less informative visual tokens. A common strategy is to leverage attention scores to identify tokens for removal. For example, PruMerge (Shang et al., 2024b) clusters and merges tokens in the vision encoder based on attention sparsity, while FastV (Chen et al., 2024a) utilizes attention weights from the second layer of the LLM to guide pruning. SparseVLM (Zhang et al., 2024b) employs cross-modal attention for text-conditioned token selection, and VisionZip (Yang et al., 2025) compresses visual inputs via CLS token attention in the final vision encoder layer. FlowCut (Tong et al., 2025) identifies redundancy from an information flow perspective by analyzing attention propagation. LVPruning (Sun et al., 2025) uses cross-attention to assess vision token importance via interaction with language tokens, guiding token pruning. Other methods adopt calibration-based strategies, where pruning ratios and depths are determined by evaluating model behavior on a held-out set. FitPrune (Ye et al., 2025b) compares attention distributions before and after pruning to inform token selection, while VTW (Lin et al., 2025a) shows that tokens can be safely dropped after specific layers, guided by calibration. In addition, approaches like DivPrune (Alvar et al., 2025) addresses token pruning by maximizing the minimum distance between tokens. Moreover, a concurrent work (Zhang et al., 2025) reformulates the token pruning problem using determinantal point processes to achieve effective pruning.

3 Token Redundancy Across Modalities

Previous studies (Shang et al., 2024a; Alvar et al., 2025; Li et al., 2025a) primarily decide to prune visual tokens based on the discrepancy in the number of visual and textual tokens. However, these works do not thoroughly analyze the differences in redundancy across modalities. In this work, we provide a more detailed pruning guidance by examining the redundancy of visual and textual tokens from **three perspectives: attention mechanisms, semantic distribution and information repetition**. Our findings reveal that visual tokens exhibit substantially higher redundancy than textual tokens. Motivated by this observation, we retain all textual tokens during the pruning process and apply pruning exclusively to visual tokens. The detailed analysis can be found in Appendix A.

4 Methodology

4.1 Problem Formulation

In this section, we propose a visual token pruning strategy that explicitly preserves cross-modal alignment while promoting intra-modal information diversity. Given an input image, we extract a sequence of visual tokens $\mathcal{X}^v = \{x_1^v, \ldots, x_N^v\}$, and embed the accompanying text into a sequence of textual tokens $\mathcal{X}^t = \{x_1^t, \ldots, x_M^t\}$, where $x_i^v, x_j^t \in R^d$.

Our pruning pipeline can be roughly divided into two stages. In the first stage, we select a subset $\mathcal{X}^{v(1)} \subset \mathcal{X}^v$ with size $\left|\mathcal{X}^{v(1)}\right| = N_1$, where $N_1 < N$, by retaining tokens with the highest semantic alignment to the text modality, thereby ensuring the preservation of tokens with strong cross-modal correlations. In the second stage, we further prune $\mathcal{X}^{v(1)}$ to obtain a more compact subset $\mathcal{X}^{v(2)} \subset \mathcal{X}^{v(1)}$ with size $\left|\mathcal{X}^{v(2)}\right| = N_2$, where $N_2 < N_1$, by maximizing the expected pairwise distance in the embedding space. This promotes non-redundant and information-rich representations while reducing computational overhead. We provide the experimental analysis of the stage order and the used metrics in Section 5.5 and Appendix D. Figure 1 illustrates the overview of our method.

4.2 Cross-Modal Alignment Aware Token Filtering

During training, multimodal models primarily optimize cross-entropy loss for text generation, which often leads to an implicit weakening of cross-modal

alignment (Covert et al., 2024). Pruning as a post-processing step for the model is crucial for cross-modal alignment. To preserve cross-modal alignment during visual token pruning, we introduce a mutual information based criterion that selects a subset of visual tokens $\mathcal{X}^{v(1)}$ which maximizes the expected mutual information with the textual tokens. Our selection objective can be formulated as follows:

$$\mathcal{X}^{v(1)} = \underset{\substack{\mathcal{X}^{v'} \subset \mathcal{X}^v, \\ |\mathcal{X}^{v'}| = N_1}}{\arg\max} E_{x^t \in \mathcal{X}^t} \left[I(\mathcal{X}^{v'}; x^t) \right] \quad (1)$$

Here, $\mathcal{X}^{v'}$ serves as a temporary variable. $I(\mathcal{X}^{v'}; x^t)$ denotes the average mutual information between the textual token x^t and all visual tokens in the subset $\mathcal{X}^{v'}$. Furthermore, we define the alignment score α_i of each visual token x_i^v based on its mutual information with the textual token set $\mathcal{X}^t$ as follows:

$$\alpha_i = I(x_i^v; \mathcal{X}^t) = \frac{1}{M} \sum_{j=1}^{M} I(x_i^v; x_j^t) \quad (2)$$

This score reflect the average semantic alignment strength of visual tokens within the shared embedding space. Intuitively, tokens with higher alignment scores are retained, while those with lower semantic alignment are pruned.

Although mutual information provides a theoretically grounded measure of cross-modal alignment, its precise estimation in high dimensional spaces often requires costly density modeling. We approximate mutual information using the L_2 norm, which reduces computational overhead without compromising performance. It is worth noting that our approximation is grounded in mathematical theory and does not rely on any specific model architecture. **Additionally, we provide theoretical justification and experimental analysis to explain why the L_2 norm can be used for approximation (see Appendix B and D).** Therefore, we can compute α_i using the following formula:

$$\alpha_i = -\frac{1}{M} \sum_{j=1}^{M} \|x_i^v - x_j^t\|_2 \quad (3)$$

We then construct the set $\mathcal{X}^{v(1)}$ by selecting the N_1 visual tokens with the highest alignment scores, which can be represented as follows:

$$\mathcal{X}^{v(1)} = \{x_i^v \in \mathcal{X}^v \mid \alpha_i \in top_{N_1}(\alpha_1, \alpha_2, \ldots, \alpha_N)\} \quad (4)$$

This selection effectively filters out semantically misaligned visual tokens while retaining those with the strongest cross-modal correlations, thereby enhancing alignment quality and reducing computational overhead. Even for semantically sparse queries (e.g., image captioning on the COCO dataset), our method consistently achieves strong performance.

Superiority over Other Works Some works prune visual tokens based on their attention scores with textual tokens. However, these approaches are susceptible to positional bias, often retaining spatially adjacent tokens that are highly redundant. Although both our method and these approaches leverage interactions between text and vision, our method emphasizes alignment in the semantic space rather than relying on positional correlations. Moreover, we do not rely on an additional CLIP model to remap the embeddings, as this would alter the vector space and introduce extra computational overhead. We provide a detailed comparison under different metrics in Appendix D, where the results show that the L_2 norm achieves better performance compared to other measures.

4.3 Greedy Intra-Modal Representational Maximization (RepMax)

After obtaining the visual token set $\mathcal{X}^{v(1)}$, we select a refined subset $\mathcal{X}^{v(2)}$ by explicitly maximizing intra-modal information diversity. The core idea is to retain a set of visual tokens that are semantically diverse and non-redundant. To achieve this, we maximize the expected pairwise distance among the selected tokens. Specifically, we define the distance between two visual tokens x_i^v and x_j^v using cosine dissimilarity. Based on this, we formulate the objective for subset selection as follows:

$$D(x_i^v, x_j^v) = 1 - \frac{(x_i^v)^\top x_j^v}{\|x_i^v\|_2 \cdot \|x_j^v\|_2} \quad (5)$$

$$\mathcal{X}^{v(2)} = \underset{\substack{\widetilde{\mathcal{X}} \subset \mathcal{X}^{v(1)}, \\ |\widetilde{\mathcal{X}}| = N_2}}{\arg\max} E_{x_i^v, x_j^v \in \widetilde{\mathcal{X}}, i \neq j} \left[D(x_i^v, x_j^v) \right] \quad (6)$$

This objective encourages the model to retain visual tokens that capture complementary aspects of the input, thereby enhancing the representational richness of the pruned token set. To address the problem, we introduce a binary selection vector $\boldsymbol{\gamma} \in \{0, 1\}^{N_1}$. The objective can be expressed as

follows:

$$\max_{\boldsymbol{\gamma}\in\{0,1\}^{N_1}} \sum_{\substack{i=1,j=1\\ i\neq j}}^{N_1} \gamma_i\gamma_j \cdot D(x_i^v, x_j^v), \quad \sum_{i=1}^{N_1} \gamma_i = N_2 \quad (7)$$

The above problem is a combinatorial optimization task, which is known to be NP Hard. To address this, we adopt an efficient greedy algorithm for approximate optimization. The core idea is to iteratively select the next visual token that is farthest from the current selected subset in the embedding space, thereby progressively enhancing overall information diversity. Our experimental results further validate the effectiveness of the greedy algorithm.

At the initial stage of the greedy algorithm, we quantify pairwise relationships among visual tokens by constructing a cosine similarity matrix $\mathcal{C} \in R^{N_1\times N_1}$. Each element $\mathcal{C}_{ij}$ of the matrix $\mathcal{C}$ represents the similarity between tokens x_i^v and x_j^v. Based on this matrix, we compute the average similarity of each token with all others. The token with the lowest average similarity is chosen as the initial seed to initialize the selected set for the first iteration ($\mathcal{S}^{(1)}$), and the remaining tokens constitute the initial remaining set ($\mathcal{R}^{(1)}$). This can be formulated as follows:

$$\mathcal{C}_{ij} = \frac{(x_i^v)^\top x_j^v}{\|x_i^v\|_2 \cdot \|x_j^v\|_2} \quad (8)$$

$$c_i^{(1)} = \frac{1}{N_1}\sum_{j=1}^{N_1} \mathcal{C}_{ij}, \quad \forall i \in [N_1] \quad (9)$$

$$s^{(1)} = \arg\min_{i\in[N_1]} c_i^{(1)} \quad (10)$$

$$\mathcal{S}^{(1)} = \{s^{(1)}\},\ \mathcal{R}^{(1)} = [N_1] \setminus \{s^{(1)}\} \quad (11)$$

where $c_i^{(1)}$ denotes the average similarity between token x_i^v and the other N_1 visual tokens. The index $s^{(1)}$ corresponds to the initial token selected as the least similar on average to all others, and $[N_1]$ denotes the set of indexs from 1 to N_1. To enable efficient incremental computation, we maintain a similarity accumulation vector $\sigma \in R^{N_1}$, where each element σ_i denotes the average similarity between token x_i^v and all tokens in the selected set. The initial similarity accumulation vector $\sigma^{(1)}$ is given by the similarity between the token indexed by $s^{(1)}$ and all other tokens, which can be represented as follows:

$$\sigma^{(1)} = \mathcal{C}_{s^{(1)}:} \quad (12)$$

where $\mathcal{C}_{s^{(1)}:}$ denotes the row vector in the matrix $\mathcal{C}$ corresponding to the index $s^{(1)}$.

At each iteration $t = 2, \ldots, N_2$, we compute the average similarity between each remaining token and all tokens in the current selected set. The remaining token with the lowest average similarity is selected in this iteration, and its index is used to update both the selected and remaining sets. Meanwhile, the total similarity accumulation vector is incrementally updated to incorporate the newly selected token. This process is formalized as follows:

$$c_i^{(t)} = \frac{1}{t-1} \cdot \sigma_i^{(t-1)}, \quad \forall i \in \mathcal{R}^{(t-1)} \quad (13)$$

$$s^{(t)} = \arg\min_{i\in\mathcal{R}^{(t-1)}} c_i^{(t)} \quad (14)$$

$$\mathcal{S}^{(t)} = \mathcal{S}^{(t-1)} \cup \{s^{(t)}\},\ \mathcal{R}^{(t)} = \mathcal{R}^{(t-1)} \setminus \{s^{(t)}\} \quad (15)$$

$$\sigma^{(t)} = \sigma^{(t-1)} + \mathcal{C}_{s^{(t)}:} \quad (16)$$

Here, $\mathcal{S}^{(t)}$ and $\mathcal{R}^{(t)}$ denote the selected and remaining sets after the t-th iteration, respectively. $c_i^{(t)}$ represents the average similarity between the token x_i^v in the remaining set and all tokens in the selected set at iteration t. The index $s^{(t)}$ corresponds to the token selected during the t-th iteration, and $\mathcal{C}_{s^{(t)}:}$ denotes the row vector in the matrix $\mathcal{C}$ corresponding to the index $s^{(t)}$. The similarity accumulation vector $\sigma^{(t)}$ is updated using the similarity vector $\mathcal{C}_{s^{(t)}:}$ of the newly selected token.

This process is repeated N_2 times until N_2 tokens have been selected. The resulting set $\mathcal{X}^{v(2)} \subset \mathcal{X}^{v(1)}$ forms a maximally dispersed subset of visual tokens with minimal internal redundancy.

Superiority over Other Works We compare our method with DivPrune, which performs pruning by maximizing the minimum distance. In our experiments, we observe that outliers in the token embeddings can significantly affect DivPrune, preventing it from achieving optimal performance. In contrast, our method optimizes the expected pairwise distance, which effectively mitigates the influence of outliers, and employs a designed greedy algorithm to efficiently tackle the associated NP hard problem. The results of our comparative experiments demonstrate a substantial performance improvement over DivPrune. Furthermore, our method maintains strong performance when compared to an extended variant (DivPrune*) implemented by our method (see Section 5.4).

Method	ChartQA	COCO	MMB	MME	MMU	NoCaps	OCRB	VQA_{OK}	POPE	VQA_{TEXT}
Upper Bound, 576 Tokens (100%)										
Dense	18.20	1.10	64.09	1508.24	36.33	1.06	31.30	53.44	93.86	46.11
Retain 192 Tokens (↓ 66.7%)										
ToMe	15.45	0.08	58.64	1262.74	31.49	0.08	29.19	46.88	77.64	39.60
FastV	15.24	0.08	63.06	1302.33	31.26	0.09	29.53	47.24	69.49	38.43
SpVLM	17.17	1.05	61.62	1390.39	35.55	1.00	30.16	50.47	89.66	42.00
PDrop	17.06	1.06	60.41	1426.75	34.42	0.99	30.35	50.83	88.26	41.78
VisionZip	17.48	1.06	62.92	1440.02	35.88	1.01	30.41	51.14	92.20	43.01
DivPrune	17.40	1.06	62.29	1436.44	35.78	1.01	30.40	51.55	91.37	43.24
Ours	**17.80**	**1.08**	**63.75**	**1467.30**	**36.22**	**1.03**	**31.00**	**52.29**	**93.68**	**45.10**
Retain 128 Tokens (↓ 77.8%)										
ToMe	14.13	0.06	57.27	1066.04	30.69	0.06	24.93	43.24	68.92	37.90
FastV	13.86	0.07	**62.89**	1182.73	31.71	0.08	26.24	44.56	65.41	35.88
SpVLM	16.56	1.04	60.08	1346.25	33.85	1.01	28.06	48.35	88.35	40.51
PDrop	16.35	1.03	59.97	1320.85	34.76	1.01	28.81	49.85	90.32	41.30
VisionZip	17.00	1.05	61.94	1393.33	36.10	1.02	28.88	49.66	93.75	42.71
DivPrune	16.96	1.04	61.77	1396.25	36.22	**1.04**	29.00	50.02	91.31	41.66
Ours	**17.80**	**1.06**	62.80	**1397.22**	**36.44**	1.02	**29.50**	**51.12**	**94.29**	**44.28**
Retain 64 Tokens (↓ 88.9%)										
ToMe	12.49	0.05	49.97	909.99	27.46	0.06	20.07	39.47	62.85	34.59
FastV	12.03	0.05	58.94	1004.35	29.40	0.06	22.04	41.65	57.46	32.81
SpVLM	14.92	0.99	58.03	1203.46	26.55	0.95	25.81	45.14	89.91	37.51
PDrop	15.14	0.99	45.42	1248.24	34.98	0.94	27.00	44.09	66.92	33.81
VisionZip	15.74	1.00	59.91	1348.03	35.80	0.95	28.50	48.85	92.85	41.31
DivPrune	15.84	0.99	59.28	1348.99	35.89	0.94	27.60	48.36	92.18	39.22
Ours	**16.36**	**1.02**	**61.34**	**1359.12**	**36.11**	**0.98**	**29.00**	**49.37**	**95.01**	**41.45**

Table 1: Comparison of visual token pruning methods on LLaVA-1.5-7B across multiple benchmarks under varying token retention ratios. Best results are highlighted in bold.

5 Experiments

5.1 Experimental Setup

Models and Baselines. We evaluate the performance of our method across several representative LMMs, including LLaVA-1.5-7B (Liu et al., 2023), LLaVA-1.5-13B (Liu et al., 2023), LLaVA-NEXT-7B (Liu et al., 2024b), LLaVA-NEXT-Video-7B (Zhang et al., 2024c) and Qwen2-VL-7B (Wang et al., 2024). These models span a range of parameter scales and encompass both image and video understanding tasks. We evaluate our method alongside several widely adopted baselines, including ToMe (Bolya et al., 2023), FastV (Chen et al., 2024a), SparseVLM (SpVLM) (Zhang et al., 2024b), PDrop (Xing et al., 2024), VisionZip (Yang et al., 2025) and DivPrune (Alvar et al., 2025). We additionally include the VTW (Lin et al., 2025b) method in our efficiency evaluation. All baselines are assessed under consistent experimental settings.

Datasets and Metrics. We evaluate our method on diverse multimodal benchmarks covering various tasks. For ChartQA (Masry et al., 2022), we report the relaxed score. Image captioning quality and diversity are measured by CIDEr on COCO (Sharma et al., 2018) and NoCaps (Agrawal et al., 2019). $MMBench_{EN}$ (MMB) (Liu et al., 2024d) uses a GPT-based score. Perceptual understanding is assessed via the perception score on MME (Fu et al., 2023). Accuracy evaluates general reasoning on MMU (Zheng et al., 2025) and text recognition on OCRBench (Liu et al., 2024e). Exact match accuracy measures question answering on VQA_{OK} (Marino et al., 2019) and VQA_{TEXT} (Singh et al., 2019). POPE (Li et al., 2023b) is evaluated with precision for positional understanding. This comprehensive protocol ensures thorough assessment of generalization across multimodal tasks.

Implementation Details. Our experiments are conducted using the PyTorch framework (Paszke et al., 2019) and the Hugging Face Transformers library (Wolf, 2020). We utilize an NVIDIA H100 GPU with 80GB of memory. We set the parameter $N_1 = \lfloor 0.8N \rceil$ by default. In both the ablation study and case study, we fix the final number of retained tokens to 64. The temperature, prompt

and preprocessing in our experiments all follow the default settings of LMMs-Eval (Bo Li* and Liu, 2024).

Method	ChartQA	COCO	MMB	NoCaps	OCRB	POPE
	Upper Bound, 576 Tokens (100%)					
Dense	18.20	1.16	68.73	1.09	33.60	94.44
	Retain 144 Tokens (↓ 75.0%)					
VisionZip	**17.44**	**1.10**	65.76	1.03	31.19	95.54
DivPrune	17.28	1.09	66.41	1.02	31.10	94.63
Ours	**17.44**	**1.10**	**67.35**	**1.04**	**33.00**	**96.69**
	Retain 128 Tokens (↓ 77.8%)					
VisionZip	17.09	**1.09**	66.02	1.02	31.01	95.04
DivPrune	17.08	**1.09**	66.07	1.01	30.90	94.76
Ours	**17.88**	**1.09**	**66.84**	**1.04**	**33.10**	**96.55**
	Retain 64 Tokens (↓ 88.9%)					
VisionZip	16.29	1.04	64.44	0.97	29.77	95.03
DivPrune	16.32	1.03	64.52	0.97	30.10	95.84
Ours	**16.56**	**1.05**	**64.95**	**0.99**	**30.60**	**97.63**

Table 2: Comparison between our visual token pruning method and DivPrune on LLaVA-1.5-13B across multiple benchmarks under varying token retention settings.

5.2 Main Results

We conduct a comprehensive evaluation of our visual token pruning approach against several baselines. The upper bound refers to the maximum number of visual tokens. As shown in Table 1, our method consistently outperforms prior works across a range of token retention ratios on LLaVA-1.5-7B. At a moderate retention of 192 tokens, our approach achieves an MME perception score of 1467.30, with only a 2.7% drop relative to the dense model, compared to 13.7% and 16.2% drops for FastV and ToMe, respectively. Compared with the strongest baseline, DivPrune, our method improves MMB accuracy from 62.28 to 63.75, a relative gain of 2.4%. To further assess the scalability of our approach, we extend our evaluation to LLaVA-1.5-13B and conduct a comparative analysis against VisionZip and DivPrune under varying visual token retention settings. As presented in Table 2, our method consistently surpasses others, demonstrating its effectiveness at higher model scale. These results highlight the generalization ability of our pruning strategy in large scale settings.

In addition, we conduct experiments on LLaVA-NEXT-7B using a larger pool of visual tokens across different retention settings. As shown in Table 3, our method consistently achieves the strongest performance across a range of compression levels. Furthermore, our experiments on LLaVA-NEXT-Video-7B and Qwen2-VL-7B also demonstrate the effectiveness of the proposed method (see Appendix C).

Method	ChartQA	COCO	MME	NoCaps	OCRB	POPE
	Upper Bound, 2880 Tokens (100%)					
Dense	54.88	1.00	1519.30	0.88	52.50	95.71
	Retain 720 Tokens (↓ 75.0%)					
ToMe	36.57	0.08	1294.12	0.09	42.25	81.58
FastV	36.09	0.09	1334.69	0.10	42.74	73.01
SpVLM	40.66	0.95	1424.94	0.82	43.65	94.20
PDrop	40.39	0.95	1462.20	0.83	43.93	92.73
VisionZip	40.91	0.95	1480.45	0.84	44.07	96.27
DivPrune	41.20	0.95	1472.13	0.83	44.00	96.00
Ours	**44.92**	**1.00**	**1503.24**	**0.86**	**46.40**	**96.44**
	Retain 640 Tokens (↓ 77.8%)					
ToMe	32.43	0.08	1133.13	0.08	37.05	72.73
FastV	31.82	0.08	1257.15	0.09	39.00	69.02
SpVLM	37.99	0.94	1430.96	0.83	41.71	93.22
PDrop	37.53	0.94	1403.96	0.82	42.82	95.31
VisionZip	39.01	0.96	1475.23	0.83	42.87	96.67
DivPrune	38.92	0.96	1484.12	0.83	43.10	96.35
Ours	**43.72**	**0.99**	**1492.60**	**0.99**	**45.10**	**96.76**
	Retain 320 Tokens (↓ 88.9%)					
ToMe	25.27	0.08	976.13	0.08	25.23	66.17
FastV	24.34	0.08	1077.35	0.08	27.71	60.50
SpVLM	30.18	0.92	1290.93	0.80	32.45	94.66
PDrop	30.62	0.92	1338.97	0.80	33.95	70.46
VisionZip	32.14	0.94	1445.85	0.81	35.35	95.71
DivPrune	32.04	0.93	1447.05	0.80	34.70	**97.05**
Ours	**35.64**	**0.96**	**1452.37**	**0.82**	**37.90**	96.47

Table 3: Performance comparison of our visual token pruning method with other baselines on LLaVA-NEXT-7B across multiple benchmarks under varying token retention settings (retaining 720, 640, and 320 tokens).

5.3 Efficiency Analysis (Accounting for Pruning Overhead)

In this section, we evaluate the inference efficiency of our method on LLaVA-NEXT-7B using the MME dataset. We repeat the process three times and report the average result. For a fair comparison, the number of decoding steps for each method is fixed to the minimum decoding steps required among all methods. As shown in Table 4, we report the average inference time (time to generate the complete output) per sample with a batch size of 1 and 288 visual tokens retained. Compared to the original dense model (15.53 GB, 235.2 ms), all pruning methods significantly reduce memory usage and cut inference latency by more than half. Our method achieves an inference time of 100.13 ms, representing a 57% reduction compared to the dense baseline. When compared with other pruning baselines, our approach achieves lower latency than VTW and FastV. Although our latency is slightly higher than DivPrune, this small gap is negligible given the superior task performance consistently achieved under the same token budget. It is worth noting that pruning is performed during inference; therefore, the reported inference latency includes the overhead introduced by the pruning process. These results demonstrate that our pruning strategy

Method	Memory (G)	Latency (ms)
Dense	15.53	235.2
VTW	13.63	103.71
FastV	13.63	108.29
DivPrune	13.63	99.53
Ours	13.63	100.13

Table 4: Comparison of peak GPU memory usage and average per-sample latency for our pruning method and baselines on LLaVA-NEXT-7B, evaluated on the MME dataset with a batch size of 1 and 288 retained tokens.

provides substantial improvements in memory and latency efficiency.

5.4 Comparison with our DivPrune variants (DivPrune*)

We further extend the DivPrune approach to enable a more comprehensive comparison. Concretely, we integrate our cross-modal pruning strategy with DivPrune under its default configuration, forming a new variant denoted as DivPrune*. We evaluate this variant on LLaVA-1.5-7B while retaining 64 visual tokens. As shown in Table 5, our method still surpasses DivPrune*. Combined with the earlier results, we also observe that DivPrune* achieves noticeable gains over the original DivPrune, highlighting the crucial role of cross-modal alignment in boosting performance. Moreover, these findings demonstrate that our greedy intra-modal expectation representation maximization strategy is more effective than the minimum distance maximization employed in DivPrune.

Method	ChartQA	COCO	MMB	MME	MMU	NoCaps
DivPrune	15.84	0.99	59.28	1348.99	35.89	0.94
DivPrune*	16.02	1.01	60.58	1355.28	36.02	0.96
Ours	16.36	1.02	61.34	1359.12	36.11	0.98

Table 5: Comparison of DivPrune variants (DivPrune*) with our method on LLaVA-1.5-7B across multiple benchmarks.

5.5 Ablation Study

To further validate the effectiveness of our approach, we conduct ablation studies with three different configurations. $ablation_1$ swaps the order of the two pruning stages by applying intra-modal representation maximization before cross-modal alignment. $ablation_2$ removes intra-modal pruning and only retains cross-modal alignment based pruning. $ablation_3$ eliminates cross-modal alignment pruning and applies only intra-modal representation maximization. As shown in Table 6, the performance consistently drops once any component is removed, indicating the necessity of each part of our framework. Moreover, using only cross-modal pruning or altering the pruning order leads to more severe degradation. We conjecture that this is because text-related visual information is often concentrated in local regions, where tokens align with the text but may carry highly similar semantics. Additional experiments in Section E further demonstrate that with properly chosen pruning parameters for each stage, our method achieves strong performance.

Method	ChartQA	COCO	MMB	MME	MMU	NoCaps
$ablation_1$	13.12	0.83	50.00	1114.77	34.56	0.76
$ablation_2$	13.08	0.83	49.57	1108.64	34.33	0.75
$ablation_3$	16.16	1.00	59.97	1311.46	36.02	0.97
Ours	16.36	1.02	61.34	1359.12	36.11	0.98

Table 6: Ablation study of different pruning strategies on LLaVA-1.5-7B across multiple benchmarks.

5.6 Additional Results

We conduct a **Hyperparameter Analysis**, such as varying N_1, and observe consistently strong performance across different settings, demonstrating the robustness of our method (Appendix E). Moreover, we perform a **Case Study** to show that the pruned model retains its ability to generate high quality content (Appendix F).

6 Conclusion

In this paper, we analyze the redundancy discrepancy between visual and textual tokens in LMMs and propose a pruning strategy that operates exclusively on visual tokens. Our method explicitly preserves cross-modal alignment and intra-modal informational diversity. Specifically, we leverage mutual information to eliminate visual tokens that are semantically misaligned with textual inputs, ensuring the consistency across modalities. To further enhance the representational quality of the retained tokens, we maximize their expected pairwise distances in the embedding space via an efficient greedy algorithm. Extensive experiments across diverse models and benchmarks demonstrate the effectiveness of our approach, alongside significant improvements in memory consumption and inference latency. We plan to extend our approach to more complex or non-visual modalities in future work, further demonstrating its broad applicability.

Limitations

In this work, we conduct extensive experiments to evaluate the effectiveness of our multimodal model token pruning method. The results demonstrate that our approach achieves competitive performance compared to the baselines. However, due to computational constraints, we have not yet been able to evaluate it on larger scale models, such as those with 70 billion parameters. Exploring the scalability of our method to such large models constitutes an important direction for future work.

References

Lada A Adamic and Bernardo A Huberman. 2002. Zipf's law and the internet. *Glottometrics*, 3(1):143–150.

Harsh Agrawal, Karan Desai, Yufei Wang, Xinlei Chen, Rishabh Jain, Mark Johnson, Dhruv Batra, Devi Parikh, Stefan Lee, and Peter Anderson. 2019. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 8948–8957.

Saeed Ranjbar Alvar, Gursimran Singh, Mohammad Akbari, and Yong Zhang. 2025. Divprune: Diversity-based visual token pruning for large multimodal models. *arXiv preprint arXiv:2503.02175*.

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*.

Kaichen Zhang* Fanyi Pu* Xinrun Du Yuhao Dong Haotian Liu Yuanhan Zhang Ge Zhang Chunyuan Li Bo Li*, Peiyuan Zhang* and Ziwei Liu. 2024. Lmms-eval: Accelerating the development of large multimoal models.

Daniel Bolya, Cheng-Yang Fu, Xiaoliang Dai, Peizhao Zhang, Christoph Feichtenhofer, and Judy Hoffman. 2023. Token merging: Your vit but faster. In *Proceedings of the International Conference on Learning Representations*.

T Brown, B Mann, N Ryder, M Subbiah, JD Kaplan, P Dhariwal, A Neelakantan, P Shyam, G Sastry, A Askell, and 1 others. 2020. Language models are few-shot learners advances in neural information processing systems 33.

Mu Cai, Jianwei Yang, Jianfeng Gao, and Yong Jae Lee. 2024. Matryoshka multimodal models. In *Workshop on Video-Language Models@ NeurIPS 2024*.

Yevgen Chebotar, Quan Vuong, Karol Hausman, Fei Xia, Yao Lu, Alex Irpan, Aviral Kumar, Tianhe Yu, Alexander Herzog, Karl Pertsch, and 1 others. 2023. Q-transformer: Scalable offline reinforcement learning via autoregressive q-functions. In *Conference on Robot Learning*, pages 3909–3928. PMLR.

Kevin Chen, Jiayuan Li, Yang Li, and Zhen Li. 2023. A survey on efficient transformer models. *arXiv preprint*.

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.

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

Wei-Lin Chiang, Zhuohan Li, Ziqing Lin, Ying Sheng, Zhanghao Wu, Hao Zhang, Lianmin Zheng, Siyuan Zhuang, Yonghao Zhuang, Joseph E Gonzalez, and 1 others. 2023. Vicuna: An open-source chatbot impressing gpt-4 with 90%* chatgpt quality. *See https://vicuna. lmsys. org (accessed 14 April 2023)*, 2(3):6.

Ian Covert, Tony Sun, James Zou, and Tatsunori Hashimoto. 2024. Locality alignment improves vision-language models. *arXiv preprint arXiv:2410.11087*.

Hugging Face. 2024. *Mastering Transformers: The Journey from BERT to Large Language Models and Stable Diffusion*. Open Library.

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.

Julia Hirschberg and Christopher D Manning. 2015. Advances in natural language processing. *Science*, 349(6245):261–266.

E. Huang and 1 others. 2023. Evaluating large language models in complex scenarios. *Journal of Computational Linguistics*.

Kai Huang, Hao Zou, Ye Xi, BoChen Wang, Zhen Xie, and Liang Yu. 2024. Ivtp: Instruction-guided visual token pruning for large vision-language models. In *European Conference on Computer Vision*, pages 214–230. Springer.

Feyza Duman Keles, Pruthuvi Mahesakya Wijewardena, and Chinmay Hegde. 2023. On the computational complexity of self-attention. In *International conference on algorithmic learning theory*, pages 597–619. PMLR.

Brett Koonce. 2021. Resnet 50. In *Convolutional neural networks with swift for tensorflow: image recognition and dataset categorization*, pages 63–72. Springer.

Jaewoo Lee, Keyang Xuan, Chanakya Ekbote, Sandeep Polisetty, Yi R Fung, and Paul Pu Liang. 2025. Tamp: Token-adaptive layerwise pruning in multimodal large language models. *arXiv preprint arXiv:2504.09897*.

C. Li and 1 others. 2023a. Fine-tuning techniques for efficient model adaptation. *AI Research Journal*.

Duo Li, Zuhao Yang, and Shijian Lu. 2025a. Todre: Visual token pruning via diversity and task awareness for efficient large vision-language models. *arXiv preprint arXiv:2505.18757*.

Mingxiao Li, Na Su, Fang Qu, Zhizhou Zhong, Ziyang Chen, Yuan Li, Zhaopeng Tu, and Xiaolong Li. 2025b. Vista: Enhancing vision-text alignment in mllms via cross-modal mutual information maximization. *arXiv preprint arXiv:2505.10917*.

Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. 2025c. Tokenpacker: Efficient visual projector for multimodal llm. *International Journal of Computer Vision*, pages 1–19.

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

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

Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. 2025a. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 5334–5342.

Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. 2025b. Boosting multimodal large language models with visual tokens withdrawal for rapid inference. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5334–5342.

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

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

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

Shizhan Liu, Hang Yu, Cong Liao, Jianguo Li, Weiyao Lin, Alex X Liu, and Schahram Dustdar. 2022. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *# PLACEHOLDER_PARENT_METADATA_VALUE#*.

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

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, and 1 others. 2024d. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pages 216–233. Springer.

Yuliang Liu, Zhang Li, Mingxin Huang, Biao Yang, Wenwen Yu, Chunyuan Li, Xu-Cheng Yin, Cheng-Lin Liu, Lianwen Jin, and Xiang Bai. 2024e. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

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*.

Mark EJ Newman. 2005. Power laws, pareto distributions and zipf's law. *Contemporary physics*, 46(5):323–351.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, and 1 others. 2019. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32.

Fabian Pedregosa, Gael Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, and 1 others. 2011. Scikit-learn: Machine learning in python journal of machine learning research. *Journal of machine learning research*, 12:2825–2830.

Ziheng Qin, Zhaopan Xu, Yukun Zhou, Zangwei Zheng, Zebang Cheng, Hao Tang, Lei Shang, Baigui Sun, Xiaojiang Peng, Radu Timofte, and 1 others. 2024.

Dataset growth. In *European Conference on Computer Vision*, pages 58–75. Springer.

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

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

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

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.

Yizheng Sun, Yanze Xin, Hao Li, Chenghua Lin, and Riza Batista-Navarro. Velar: Vision-oriented language-attentive token reduction for multimodal large language models.

Yizheng Sun, Yanze Xin, Hao Li, Jingyuan Sun, Chenghua Lin, and Riza Batista-Navarro. 2025. Lvpruning: An effective yet simple language-guided vision token pruning approach for multi-modal large language models. *arXiv preprint arXiv:2501.13652*.

Hind Taud and Jean-Franccois Mas. 2017. Multilayer perceptron (mlp). In *Geomatic approaches for modeling land change scenarios*, pages 451–455. Springer.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, and 1 others. 2025. Kimi-vl technical report. *arXiv preprint arXiv:2504.07491*.

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 preprint arXiv:2505.19536*.

H Touvron, T Lavril, G Izacard, X Martinet, MA Lachaux, T Lacroix, B Rozière, N Goyal, E Hambro, F Azhar, and 1 others. 2023. Open and efficient foundation language models. *Preprint at arXiv. https://doi. org/10.48550/arXiv*, 2302.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint*.

F. Wang and 1 others. 2023. Practical applications of llms in specialized domains. *Specialized AI Applications*.

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*.

Thomas Wolf. 2020. Transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

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*.

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

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. 2025a. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 22128–22136.

Weihao Ye, Qiong Wu, Wenhao Lin, and Yiyi Zhou. 2025b. Fit and prune: Fast and training-free visual token pruning for multi-modal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 22128–22136.

Hongxu Yin, Arash Vahdat, Jose M Alvarez, Arun Mallya, Jan Kautz, and Pavlo Molchanov. 2022. A-vit: Adaptive tokens for efficient vision transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 10809–10818.

Li Yuan, Yunpeng Chen, Tao Wang, Weihao Yu, Yujun Shi, Zi-Hang Jiang, Francis EH Tay, Jiashi Feng, and Shuicheng Yan. 2021. Tokens-to-token vit: Training vision transformers from scratch on imagenet. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 558–567.

D. Zhang and 1 others. 2023. Parameter-efficient fine-tuning methods for llms. *Journal of Machine Learning Research*.

Qizhe Zhang, Aosong Cheng, Ming Lu, Zhiyong Zhuo, Minqi Wang, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. 2024a. [cls] attention is all you need for training-free visual token pruning: Make vlm inference faster. *arXiv preprint arXiv:2412.01818*.

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

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

Yuanhan Zhang, Bo Li, haotian Liu, Yong jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. 2024c. Llava-next: A strong zero-shot video understanding model.

Li Zheng, Hao Fei, Ting Dai, Zuquan Peng, Fei Li, Huisheng Ma, Chong Teng, and Donghong Ji. 2025. Multi-granular multimodal clue fusion for meme understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 26057–26065.

B. Zhu and 1 others. 2023. Expanding frontiers in large language models. *AI Frontier Research*.

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

A Token Redundancy Across Modalities

Previous studies (Shang et al., 2024a; Alvar et al., 2025; Li et al., 2025a) primarily decide to prune visual tokens based on the discrepancy in the number of visual and textual tokens. However, these works do not thoroughly analyze the differences in redundancy across modalities. In this work, we provide a more detailed pruning guidance by examining the redundancy of visual and textual tokens from three perspectives: attention mechanisms, semantic distribution and information repetition. Our findings reveal that visual tokens exhibit substantially higher redundancy than textual tokens. Motivated by this observation, we retain all textual tokens during the pruning process and apply pruning exclusively to visual tokens.

The attention distribution in multimodal models is highly skewed, with textual tokens receiving significantly higher weights than visual tokens (Li et al., 2025b; Lee et al., 2025). This suggests that the model primarily relies on text for semantic understanding. Visual inputs often consist of hundreds or thousands of tokens, far exceeding the number of textual tokens. Given the quadratic complexity of attention with respect to token count, this mismatch between quantity and contribution is pronounced. Despite the large number of visual tokens, their impact on the final representation is limited, resulting in poor trade offs between computational cost and semantic gain and reflecting the high redundancy of visual tokens.

Each token in text typically carries explicit semantic meaning, such as a noun, verb, or conjunction, reflecting natural language as a highly optimized discrete encoding system with dense and relatively uniform information distribution (Hirschberg and Manning, 2015). In contrast, semantic information in images is highly concentrated in a small number of salient regions, such as foreground objects, while the majority of visual tokens correspond to low-semantic areas like sky or walls, which mostly contain low-frequency textures or superficial variations and contribute little to high-level semantic understanding. Consequently, the effective semantic density of visual tokens is significantly lower than that of textual tokens.

Compared to textual tokens, visual tokens exhibit higher information redundancy due to differences in their generation processes and underlying information structures. Textual tokens, derived from natural language, follow Zipf's law (Newman,

2005; Adamic and Huberman, 2002) and represent distinct semantic units with low repetition. In contrast, visual tokens are typically generated by uniformly partitioning images or extracting low-level features without semantic organization. For example, a sky region is divided into multiple similar patches that cluster in embedding space and jointly represent a single high-level concept, causing redundancy. Visual tokenization resembles physical partitioning rather than frequency-driven semantic abstraction, forcing models to process many similar tokens and resulting in computational and representational inefficiency.

B Proof of L_2 Norm Approximation to Mutual Information

When considering two continuous random variables x_i^v and x_j^t, corresponding to two distributions, the mutual information between them is defined as follows:

$$I(x_i^v; x_j^t) = \int p(x_i^v, x_j^t) \log \frac{p(x_i^v | x_j^t)}{p(x_i^v)} \, dx_i^v dx_j^t \quad (17)$$

To render this quantity tractable, we make the following assumptions:

- The conditional distribution $p(x_i^v \mid x_j^t)$ is modeled as an *isotropic Gaussian*. We conduct extensive experiments and find that, during the fitting process, the conditional distribution exhibits per-dimension means and standard deviations concentrated around (-0.0035) and (0.7842), respectively, which substantiates the validity of our hypothesis. Then, we can derive the following expression:

$$p(x_i^v \mid x_j^t) = \mathcal{N}(x_j^t, \sigma^2 I) \quad (18)$$

  Here, I denotes the identity matrix.

- The marginal distribution $p(x_i^v)$ is generally unknown and intractable; thus, we approximate it as a constant C_2, following common practice in variational inference and contrastive learning.

Based on the above assumptions, we obtain the following expression:

$$\log p(x_i^v \mid x_j^t) = C_1 - \frac{1}{2\sigma^2} \|x_i^v - x_j^t\|_2^2 \quad (19)$$

$$I(x_i^v; x_j^t) \approx -\frac{1}{2\sigma^2} E_{p(x_i^v, x_j^t)} \left[\|x_i^v - x_j^t\|_2^2 \right] + C \quad (20)$$

Here, $C_1 = -\frac{d}{2} \log(2\pi\sigma^2)$ is a constant independent of the inputs, and we denote $C = C_1 - C_2$. Therefore, the L_2 norm provides a theoretically grounded approximation of mutual information.

C Additional Comparative Experiments

We evaluate our method on the video based multimodal model LLaVA-NEXT-Video-7B. Using the same pruning strategy as in our image understanding experiments, we conduct evaluations on the COCO dataset and report standard image captioning metrics, including BLEU-1/2/3/4, ROUGE-L and CIDEr. We compare our method against DivPrune, a really strong baseline for visual token pruning. As shown in Table 7, our approach consistently outperforms DivPrune across all evaluation metrics under various token retention settings. For instance, with 960 tokens retained, our method improves the BLEU-4 score from 28.46 to 30.41, representing a relative gain of 6.9%. These results demonstrate the robustness of our method in video scenarios.

Method	B-1	B-2	B-3	B-4	R-L	CIDEr
	Upper Bound, 2880 Tokens (100%)					
Dense	71.07	54.78	40.50	29.24	53.58	1.02
	Retain 960 Tokens (↓ 66.7%)					
DivPrune	72.51	54.97	40.00	28.46	53.79	0.99
Ours	**73.27**	**56.63**	**42.03**	**30.41**	**54.67**	**1.04**
	Retain 720 Tokens (↓ 75.0%)					
DivPrune	72.35	54.81	39.88	28.37	53.61	0.99
Ours	**72.87**	**56.34**	**41.77**	**30.23**	**54.47**	**1.04**
	Retain 540 Tokens (↓ 81.3%)					
DivPrune	71.72	54.16	39.21	27.75	53.28	0.98
Ours	**72.34**	**55.55**	**40.92**	**29.47**	**53.75**	**1.03**
	Retain 480 Tokens (↓ 83.3%)					
DivPrune	71.28	53.72	38.88	27.47	53.13	0.97
Ours	**72.44**	**55.49**	**40.74**	**29.21**	**53.80**	**1.01**

Table 7: Performance comparison of our visual token pruning method and DivPrune on LLaVA-NEXT-Video-7B, evaluated on COCO using standard captioning metrics. B denotes BLEU, and R denotes ROUGE.

To more comprehensively assess the effectiveness of our method, we further perform pruning experiments on Qwen2-VL-7B, pruning 88.9% of tokens. We test our approach against VisionZip and DivPrune across multiple datasets. As illustrated in Table 8, our method achieves superior performance, underscoring its robustness and effectiveness even under extreme pruning scenarios.

Method	MMB	MME	VQA_{TEXT}
Dense	78.26	1894	65
VisionZip	74.26	1702.97	57.18
DivPrune	73.96	1699.03	58.29
Ours	74.90	1706.74	58.43

Table 8: Performance comparison of 88.9% token pruning on Qwen2-VL-7B.

D Metric Analysis

In addition to the theoretical analysis of approximating mutual information (MI) using the L_2 norm, we conduct empirical validation on LLaVA-1.5-7B with 64 visual tokens retained. Specifically, we implement cross-modal alignment using three metrics: mutual information, its L_2 norm approximation, and cosine similarity. We use scikit-learn (Pedregosa et al., 2011) to compute mutual information, where the library employs a nonparametric k nearest neighbors approximation method, with k set to 3 by default. As shown in Table 9, the method using the L_2 norm approximation achieves performance comparable to that of exact mutual information, and consistently outperforms the one using cosine similarity. We guess that this phenomenon occurs because semantic alignment between textual and visual tokens requires not only directional consistency but also a substantial overlap in their feature distributions. We further compare performance under different intra-modal redundancy metrics, namely cosine similarity and the L_2 norm. As shown in Table 10, cosine similarity consistently yields better results. This suggests that cosine similarity is generally more effective for measuring redundancy within unimodal data, which is consistent with prior empirical findings (Qin et al., 2024) in the literature.

Metric	ChartQA	COCO	MMB	MME	MMU	NoCaps
MI	16.38	1.02	61.38	1362.88	36.12	0.98
L_2	16.36	1.02	61.34	1359.12	36.11	0.98
cos	16.26	1.01	61.15	1357.05	35.83	0.97

Table 9: Performance comparison among various cross-modal alignment metrics across multiple benchmarks.

E Hyperparameter Analysis

In this study, we conduct an analysis of the influence of the token retention parameter N_1 in the first stage cross-modal token pruning process on the overall performance of our method. As presented in Table 11, we evaluate several representative configurations on LLaVA-1.5-7B, including $N_1 = \lfloor 0.9N \rceil$, $N_1 = \lfloor 0.8N \rceil$, $N_1 = \lfloor 0.75N \rceil$, and $N_1 = \lfloor 0.7N \rceil$, where N represents the total number of visual tokens in the dense model. In all cases, the final number of retained tokens is fixed at 64. Our method consistently achieves strong and stable performance across these different settings, indicating that it is largely insensitive to the choice of N_1. These strong results highlight the robustness of our method to different hyperparameter settings.

Metric	ChartQA	COCO	MMB	MME	MMU	NoCaps
L_2	16.32	1.0	61.26	1356.91	36.02	0.98
cos	16.36	1.02	61.34	1359.12	36.11	0.98

Table 10: Performance comparison among various intra-modal redundancy metrics across multiple benchmarks.

N_1	ChartQA	COCO	MMB	MME	MMU	NoCaps
$\lfloor 0.9N \rceil$	16.48	1.02	60.99	1332.41	36.11	0.97
$\lfloor 0.8N \rceil$	16.36	1.02	61.34	1359.12	36.11	0.98
$\lfloor 0.75N \rceil$	16.28	1.02	61.08	1334.68	36.11	0.97
$\lfloor 0.7N \rceil$	16.00	1.02	61.43	1347.34	35.78	0.97

Table 11: Impact of different values of the parameter N_1 on model performance using LLaVA-1.5-7B.

F Case Study

To qualitatively assess the effectiveness of our pruning strategy, we conduct a case study using LLaVA-1.5-7B on samples from the COCO dataset. As illustrated in Figure 2, we compare the captions generated by the original dense model and our pruned model, which retains significantly fewer visual tokens. The results show that the pruned model produces captions that remain semantically coherent and fluent, closely aligning with those generated by the dense model. This demonstrates that our method effectively removes redundant visual tokens while preserving caption generation capabilities. Overall, the case study highlights that substantial token reduction can be achieved without compromising output quality, underscoring the practical utility of our approach.

Reproducibility statement

We strive to ensure the reproducibility of our results. Full details are provided in the main paper and the appendix. Our implementation is built on PyTorch and standard open-source libraries. We provide key

Prompt: Provide a one-sentence caption for the provided image.

GT caption: A car is parked on dry grass with kites overhead.

Dense model: A white truck is parked in a field with a bunch of kites flying in the sky.

Ours: A white truck is parked in a field with a bunch of kites flying in the sky.

(a)

Prompt: Provide a one-sentence caption for the provided image.

GT caption: A large passenger jet sitting on top of an airport runway.

Dense model: A United airplane is sitting on the runway.

Ours: A United Airlines plane is sitting on the runway.

(b)

Prompt: Provide a one-sentence caption for the provided image.

GT caption: A bathroom with a bathtub and a sink overlooking a blue ocean.

Dense model: A bathroom with a large bathtub and a window.

Ours: A large bathtub is situated in a bathroom with a window above it.

(c)

Figure 2: Case study comparing captions generated by dense and pruned models (LLaVA-1.5-7B) on the COCO dataset, demonstrating output consistency despite substantial token reduction.

code implementations to facilitate reproducibility and further research.