CrisPrune: Combining Contextual Relevance and Intrinsic Saliency for Efficient Visual Token Pruning in MLLMs

Ziniu Liu[1,2] **Shuheng Zhou**[2] **Mingqing Liu**[1*] **Hao Deng**[1] **Huijia Zhu**[2*]
[1]Tongji University [2]Ant Group
clare@tongji.edu.cn liuziniu.lzn@antgroup.com

Abstract

Visual token pruning has emerged as a pivotal strategy to alleviate the computational bottleneck in Multimodal Large Language Models (MLLMs), yet it frequently compromises the integrity of visual understanding in pursuit of efficiency. Existing methods face a fundamental tension: vision-centric approaches are susceptible to the attention sink phenomenon and operate in a query-agnostic manner, whereas text-guided methods often create an overly narrow focus, discarding essential background context and failing on ambiguous queries. In this paper, we propose CrisPrune, a training-free and model-agnostic method that reconciles efficiency with understanding by integrating visual saliency and text relevance. Specifically, we introduce intrinsic visual saliency with robust normalization to identify information-rich regions characterized by significant visual features. Simultaneously, we design dual-source text relevance to synergize explicit instruction alignment with implicit scene priors. Finally, we reformulate the selection process using a Determinantal Point Process (DPP) to balance token quality and spatial diversity. Extensive experiments demonstrate that CrisPrune significantly outperforms state-of-the-art methods. On LLaVA-NeXT, it achieves a 13× decrease in FLOPs while maintaining 97% of the original performance with 94.4% of visual tokens pruned, effectively bridging the gap between efficiency and holistic understanding.

1 Introduction

Large Language Models (LLMs)(Achiam et al., 2023; Touvron et al., 2023; Bai et al., 2023a) have achieved remarkable success. Leveraging their powerful reasoning capabilities, MLLMs(Bai et al., 2023b; Team et al., 2023; Liu et al., 2023; Chen et al., 2024b) extend this intelligence to diverse modalities. To process visual inputs, MLLMs typically encode images into sequences of tokens, analogous to textual inputs. For demanding tasks such as complex visual question answering(Hudson and Manning, 2019), and fine-grained recognition(Horn et al., 2018), high-resolution inputs are essential to capture intricate image details, leading to thousands of visual tokens per image. However, the computational complexity of the attention mechanism scales quadratically with sequence length. This inefficiency imposes significant barriers to deploying high-resolution models(Liu et al., 2024b) and video-based models(Lin et al., 2023) in resource-constrained environments.

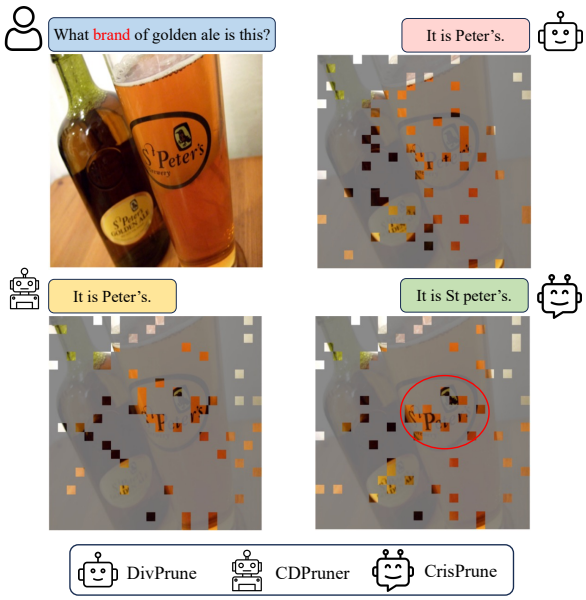

Figure 1: **Comparison between baselines and CrisPrune.** In this example, CrisPrune successfully preserves tokens related to crucial details such as the *"St. Peter's"* logo, ensuring the integrity of the visual context for reasoning.

Many efforts have been made to reduce the inference cost of MLLMs by pruning visual tokens, and existing pruning strategies fall into two main paradigms. Vision-centric methods estimate importance via raw attention scores from the vision encoder(Yang et al., 2025; Liu et al., 2025).

*Corresponding author

Findings of the Association for Computational Linguistics: ACL 2026, pages 13546–13564

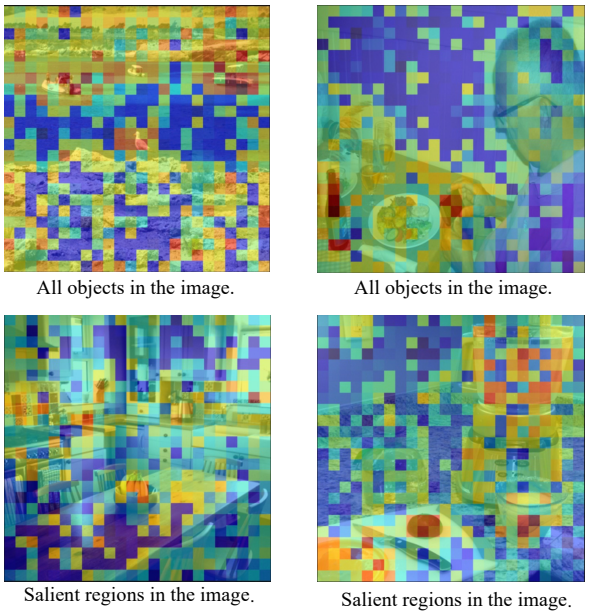

Figure 2: **Visualization of General Semantic Priors.** We compute the relevance scores using general prompts such as "*All objects in the image*" and "*Salient regions in the image*." **Red** indicates high relevance, while **blue** indicates low relevance.

However, they suffer from the attention sink phenomenon, and they operate in a query-agnostic manner. While text-interaction methods(Zhang et al., 2024; Han et al., 2025) estimate importance by incorporating semantic guidance. Early approaches like FastV(Chen et al., 2024a) leverage cross-attention within the LLM. However, they incur computational overhead and suffer from attention shift(Zhang et al., 2025a), where text tokens disproportionately attend to visual tokens that are physically closer in the sequence rather than semantically relevant. Recent text-guided approaches, such as CDPruner(Zhang et al., 2025b), calculate importance based on alignment with the input instruction. Yet, this reliance on the current instruction risks discarding essential implicit context and fails to handle generic queries. As illustrated in Figure 1, CDPruner fails to capture the prefix "St" leading to an incomplete answer. This reveals a critical gap: *how to prune redundancy without sacrificing the holistic visual context required for reasoning?*

To bridge this gap, we propose **CrisPrune**, a plug-and-play visual token pruning method that preserves holistic visual integrity by integrating intrinsic visual saliency and dual-source text relevance. Specifically, we first fully leverage intrinsic visual saliency to identify informative regions independent of textual queries. By employing a robust quantile-based normalization strategy, our method effectively retains fine-grained details while mitigating the interference of attention sinks. Simultaneously, drawing inspiration from the cognitive hierarchy in Scene Gist theory (Oliva and Torralba, 2001), we model dual-source text relevance by synergizing explicit instruction alignment with general semantic priors. As illustrated in Figure 2, these priors act as a global scanner, effectively activating background structures that might be overlooked by specific instructions. This dual-branch mechanism ensures the model remains responsive to user intent while explicitly retaining the essential environmental context, even when implicit in the query. Finally, inspired by CDPruner (Zhang et al., 2025b), we use a **Determinantal Point Process (DPP)** (Macchi, 1975) to obtain the retained tokens. This mathematical framework maximizes conditional diversity, selecting a compact subset that best represents the full visual distribution.

We evaluate our proposed approach on popular MLLMs across diverse benchmarks. Experimental results demonstrate that CrisPrune significantly outperforms existing state-of-the-art methods. For instance, on LLaVA-1.5-7B, our method retains 95% of the original performance averaged over 9 tasks while discarding 94.4% of the visual tokens. Furthermore, for models with longer visual sequences such as LLaVA-NeXT, we achieve an 11× reduction in prefilling latency, a 2.5× speedup in total inference time, and a 13× decrease in FLOPs, while retaining 97% of the original performance.

2 Related Works

2.1 Multimodal Large Language Models

Inspired by the immense success of LLMs(Achiam et al., 2023; Touvron et al., 2023) in natural language processing tasks, Multimodal Large Language Models (MLLMs)(Bai et al., 2023b; Li et al., 2023a; Liu et al., 2024a; Chen et al., 2024b) have emerged. Current mainstream MLLMs typically adopt a modular architecture, comprising three core components: a vision encoder(Zhai et al., 2023; Radford et al., 2021) that transforms visual signals into a sequence of digitized visual tokens; an LLM serving as the cognitive core, processing sequential information and generating responses; and a modality alignment module that bridges the gap by converting visual information into a format comprehensible to the LLM.

Driven by increasingly complex visual tasks

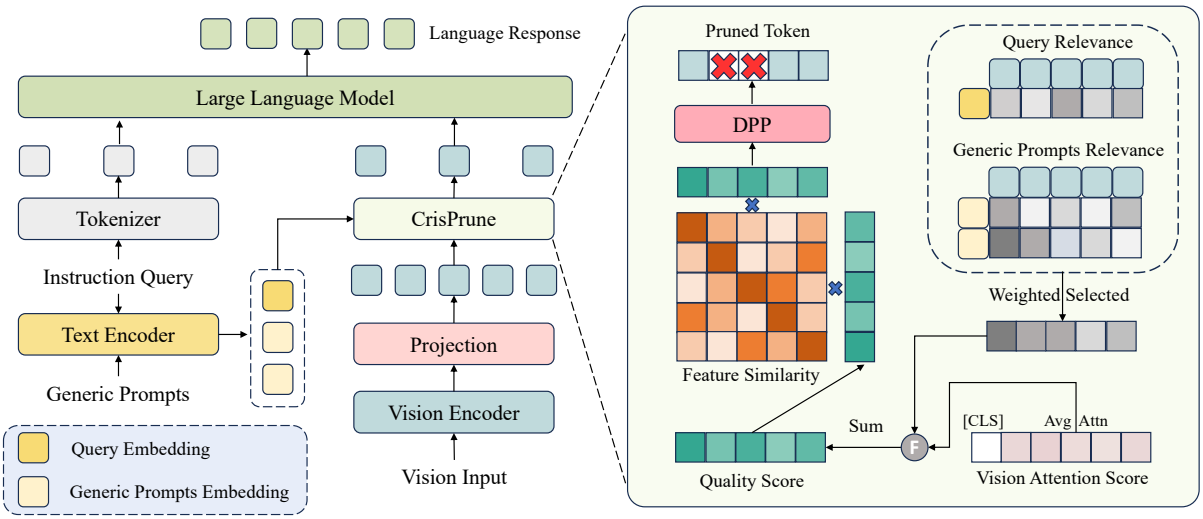

Figure 3: **The framework of CrisPrune.** Our method computes token quality scores by integrating intrinsic visual saliency and dual-source text relevance. Based on the unified quality scores, we utilize a DPP to select a representative and diverse subset of visual tokens for efficient inference.

and demands for higher precision, contemporary MLLMs continually enhance input image resolutions. This strategy enables models to capture finer visual details, leading to superior reasoning capabilities in cutting-edge models like LLaVA-NeXT(Liu et al., 2024b), Qwen2.5VL(Bai et al., 2025), and InternVL3(Zhu et al., 2025), which have achieved state-of-the-art performance. However, this strategy inherently results in a significant increase in the number of visual tokens, directly causing a drastic escalation in computational costs.

2.2 Token Reduction for MLLMs

To mitigate the high computational burden of MLLMs, a series of visual token reduction techniques have been proposed to accelerate inference. Several approaches exploit intrinsic visual properties for pruning. For instance, DART(Wen et al., 2025) identifies redundancy by removing tokens highly similar to a few pivot tokens, while DivPrune(Alvar et al., 2025) formulates the task as a Max-Min problem to maximize subset diversity. Hybrid strategies like VisPruner(Zhang et al., 2025a) and VisionZip(Yang et al., 2025) combine attention-based importance with similarity-based diversity, whereas HiPrune(Liu et al., 2025) operates via hierarchical visual attention. However, these query-agnostic methods ignore user instructions and remain susceptible to attention sinks.

Alternatively, other research incorporates textual information to guide the process. Early works primarily leverage attention scores within the LLM, such as FastV(Chen et al., 2024a) which discards tokens with low cross-attention, and VTW(Lin et al., 2025) which removes all visual tokens after a specific depth. Meanwhile, architectures like LLaMA-VID(Li et al., 2024) and LLaVA-MINI(Zhang et al., 2025c) reduce token counts by fusing multimodal information outside the backbone. More recently, approaches have focused on explicitly utilizing the correlation between visual and textual features. For example, AdaFV(Han et al., 2025) proposes an adaptive cross-modal attention blending mechanism, CATP(Li et al., 2025) implements a two-stage pruning process, GridPrune(Duan et al., 2025) introduces a zonal selection strategy utilizing text-conditional relevance and intrinsic visual saliency for spatial budget allocation, and CDPruner(Zhang et al., 2025b) employs a DPP conditioned on text relevance. Despite advancements, strategies relying on LLM attention suffer from positional bias, making it difficult to stably identify key information. Furthermore, relying exclusively on text relevance often leads to an overly narrow focus, resulting in the loss of intrinsic visual details and essential background context.

3 Method

3.1 Preliminaries

Multimodal Large Language Model Architecture. MLLMs generally consist of three components: a Vision Encoder Φ_v, a Modality Projector ϕ, and an LLM Φ_{llm}. Given an input image $\mathcal{I}$, the encoder extracts a sequence of visual tokens $\mathbf{X} \in \mathbb{R}^{N \times C}$, where N is the sequence length.

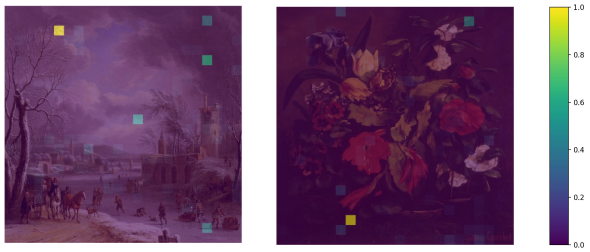

Figure 4: **Visualization of the attention sink phenomenon.** Raw attention maps show that outlier tokens (yellow) dominate the distribution, suppressing informative details (dark purple). This disparity highlights the necessity of our robust quantile-based normalization.

These tokens are projected into the embedding space by ϕ and processed by the LLM alongside text instructions to generate the response.

Visual Token Pruning. The large number of visual tokens N constitutes the primary bottleneck for inference. We aim to select a representative subset of indices $\mathcal{J} \subset \{1, \ldots, N\}$ subject to a budget constraint $|\mathcal{J}| = K$ (where $K \ll N$). Let $\mathbf{X}_{\mathcal{J}}$ denote the retained tokens. The optimization objective is to minimize the discrepancy between the model outputs generated by the full and pruned visual sequences:

$$\min_{\mathcal{J}} \mathcal{D}_{\mathrm{KL}} \left(\Phi_{llm}(\phi(\mathbf{X})) \parallel \Phi_{llm}(\phi(\mathbf{X}_{\mathcal{J}})) \right), \quad (1)$$

where $\mathcal{D}_{\mathrm{KL}}$ denotes the Kullback-Leibler divergence, measuring the information loss in the output probability distribution caused by pruning.

3.2 Intrinsic Visual Saliency

We leverage the self-attention maps from the vision encoder to derive an intrinsic saliency score. We define the raw saliency vector $\mathbf{a} \in \mathbb{R}^N$ based on the specific architecture of the vision encoder. For models equipped with a [CLS] token, we define $\mathbf{a}$ by extracting the attention weights that the [CLS] token assigns to each visual patch. Conversely, for models without a [CLS] token, we calculate a_i by averaging the attention weights received by the i-th token from all other tokens in the sequence. In both scenarios, the scalar a_i serves as a proxy for the global importance of the i-th visual token.

A direct application of min-max normalization to $\mathbf{a}$ is suboptimal due to the attention sink phenomenon (Zhang et al., 2025a) in Vision Transformers. As illustrated in Figure 4, a few specific outlier tokens accumulate disproportionately high attention weights, stretching the normalization scale and suppressing the majority of informative tokens to near-zero values.

To address this, we propose a *Robust Min-Max Normalization* strategy. We replace the absolute maximum with a upper bound v_{high} determined by a specific high quantile of the score distribution (e.g., the 98-th percentile $\tau = 0.98$). The final intrinsic visual saliency score s_i^{vis} is calculated as:

$$s_i^{vis} = \frac{\min(a_i, v_{high}) - \min(\mathbf{a})}{v_{high} - \min(\mathbf{a}) + \epsilon}, \quad (2)$$

where $\min(\mathbf{a})$ denotes the minimum scalar value in vector $\mathbf{a}$, and $\epsilon = 10^{-6}$ is a small constant added for numerical stability.

3.3 Dual-Source Text Relevance

To align the selected visual tokens with user intent, we quantify the semantic relevance between each token and the instruction. The extraction of aligned features depends on the architecture: for models with paired encoders (e.g., CLIP(Radford et al., 2021)), we employ the native text encoder; for models utilizing only a vision encoder, we take the multimodal projector's output as the visual embedding and the average of the instruction's LLM token embeddings as the text representation.

Let $\mathbf{t}_{inst}$ denote the resulting text embedding and $\mathbf{h}_i = \phi(\mathbf{x}_i)$ denote the projected visual embedding for the i-th token. We compute the instruction-specific relevance score r_i^{inst} via cosine similarity:

$$r_i^{inst} = \mathrm{sim}(\mathbf{h}_i, \mathbf{t}_{inst}) = \frac{\mathbf{h}_i \cdot \mathbf{t}_{inst}}{\|\mathbf{h}_i\| \|\mathbf{t}_{inst}\|}. \quad (3)$$

However, relying exclusively on specific instructions risks creating an overly narrow focus. This can cause the model to discard vital environmental context or salient entities not mentioned, particularly when user queries are vague (e.g., "*Describe this image*") or rely on implicit background knowledge. To mitigate this, we propose a *General Semantic Prior* to capture broad scene information. Motivated by the multiple semantic levels of visual context(Oliva and Torralba, 2001), we construct a set of generic prompts $\mathcal{T}_{gen}$ comprising "All objects", "Salient regions", and "Visible text and numbers" to cover different dimensions. These prompts are encoded using the same strategy as the instruction, and we calculate the similarity between each visual token $\mathbf{h}_i$ and these generic anchors. The general relevance is defined as the maximum similarity across these anchors:

$$r_i^{gen} = \max_{\mathbf{t} \in \mathcal{T}_{gen}} \left(\mathrm{sim}(\mathbf{h}_i, \mathbf{t}) \right). \quad (4)$$

Method	MME	TextVQA	SQA	MM-Vet	MMB-CN	GQA	VQA^{v2}	MMBench	POPE	Avg. Rel (%)
Vanilla, 576 Tokens ***(100%)***										
LLaVA-1.5-7B	1506.5	58.2	69.5	31.3	58.1	61.9	78.5	64.7	85.9	100.0%
Retain 128 Tokens (↓ 77.8%)										
FastV(ECCV24)	1368.9	56.4	69.2	27.0	55.9	54.0	71.0	63.0	68.2	92.8%
SparseVLM(ICML25)	1399.3	56.3	69.0	29.7	56.9	57.3	75.1	62.6	83.1	96.3%
PruMerge+(2024.05)	1408.1	54.0	69.1	30.4	55.8	58.2	75.0	61.8	83.1	96.8%
DART(EMNLP25)	1408.7	56.3	69.1	30.9	57.3	57.9	74.7	60.7	80.4	96.9%
DivPrune(CVPR25)	1405.1	55.9	68.6	30.6	54.8	59.4	76.0	61.5	87.0	97.5%
VisionZip(CVPR25)	1436.9	56.9	68.7	31.6	57.0	57.6	75.6	62.1	83.3	97.6%
GridPrune(2025.11)	1423.9	54.9	68.5	32.4	-	59.6	76.2	62.4	86.2	97.6%
CDPruner(NeurIPS25)	1431.4	56.2	69.0	32.8	55.0	59.9	76.6	63.1	87.7	98.2%
CrisPrune	1454.7	57.1	69.4	34.0	56.7	60.0	76.9	63.0	87.5	**99.4%**
Retain 64 Tokens (↓ 88.9%)										
FastV(ECCV24)	973.5	51.6	70.1	18.9	42.1	46.0	55.9	50.1	35.5	74.9%
SparseVLM(ICML25)	1190.4	52.1	69.2	24.4	49.6	52.0	66.9	58.3	69.7	87.1%
PruMerge+(2024.05)	1316.8	52.0	69.5	28.0	52.1	55.4	71.3	59.6	75.7	92.4%
DART(EMNLP25)	1365.1	54.7	69.3	26.5	54.0	54.7	71.3	59.5	73.8	92.6%
VisionZip(CVPR25)	1365.2	55.5	69.0	29.4	55.4	55.1	72.4	60.1	77.0	94.4%
DivPrune(CVPR25)	1334.7	54.5	68.0	28.1	52.3	57.5	74.1	60.1	85.5	94.7%
GridPrune(2025.11)	1399.2	54.3	68.2	29.3	-	58.7	75.3	62.3	85.8	95.6%
CDPruner(NeurIPS25)	1415.1	55.3	68.1	30.5	53.2	58.6	75.4	61.1	87.5	95.9%
CrisPrune	1423.8	56.5	68.4	32.1	56.0	58.9	75.7	61.9	87.5	**97.6%**
Retain 32 Tokens (↓ 94.4%)										
PruMerge+(2024.05)	1236.6	49.2	67.9	24.7	45.9	52.9	65.6	55.1	66.7	86.1%
VisionZip(CVPR25)	1251.2	53.1	69.1	25.3	50.3	51.8	67.1	57.0	69.4	88.4%
DART(EMNLP25)	1273.3	52.2	69.3	25.0	50.0	52.9	67.1	58.5	69.1	88.6%
DivPrune(CVPR25)	1284.9	52.9	68.6	26.3	49.1	54.9	71.2	57.6	81.5	91.3%
CDPruner(NeurIPS25)	1373.0	53.2	69.5	27.8	49.6	57.0	73.6	59.6	87.9	93.0%
CrisPrune	1398.5	55.0	69.3	30.7	54.0	57.2	74.1	59.7	87.6	**95.5%**

Table 1: **Performance comparison of CrisPrune on LLaVA-1.5-7B against different pruning methods across different retention ratios.** The vanilla number of visual tokens is 576. The final column (Avg. Rel) shows the average performance relative to the original model.

Finally, we fuse these two sources, employing a decay factor $\lambda \in [0, 1]$ to balance task-specific focus with broad context preservation:

$$s_i^{sem} = \max(r_i^{inst}, \lambda \cdot r_i^{gen}). \tag{5}$$

3.4 Quality-Weighted Determinantal Point Process

We synthesize the computed robust visual saliency s_i^{vis} and dual-source semantic relevance s_i^{sem} into a unified quality vector $\mathbf{q} \in \mathbb{R}^N$. Each element q_i reflects the comprehensive importance of the i-th token, derived via a combination controlled by a balance factor $\alpha \in [0, 1]$:

$$q_i = \alpha \cdot s_i^{sem} + (1 - \alpha) \cdot s_i^{vis}. \tag{6}$$

While q_i captures individual importance, direct top-k selection often yields spatially redundant tokens. To effectively mitigate this redundancy and ensure coverage diversity, we model the subset selection using a Determinantal Point Process. Let $\mathbf{S} \in \mathbb{R}^{N \times N}$ denote the base similarity matrix, where $\mathbf{S}_{ij} = \text{sim}(\mathbf{h}_i, \mathbf{h}_j)$ represents the cosine similarity between visual features.

Following (Zhang et al., 2025b), we modulate this base kernel with the quality scores to obtain a conditional kernel matrix $\mathbf{L}$:

$$\mathbf{L} = \text{diag}(\mathbf{q}) \cdot \mathbf{S} \cdot \text{diag}(\mathbf{q}). \tag{7}$$

The objective of finding the optimal subset S is equivalent to maximizing the log-determinant of the principal submatrix $\mathbf{L}_S$. This formulation explicitly decomposes the objective into quality and diversity terms:

$$\log\det(\mathbf{L}_S) = \sum_{i \in S} \log(q_i^2) + \log\det(\mathbf{S}_S). \tag{8}$$

The first term, $\sum \log(q_i^2)$, encourages the selection of tokens with high intrinsic quality, while the second term, $\log\det(\mathbf{S}_S)$, maximizes the feature diversity within the subset. Since finding the subset

Method	GQA	TextVQA	POPE	MMB-CN	VQA^{v2}	Rel
*Vanilla, 2880 Tokens (**100%**)*						
LLaVA-NeXT-7B	62.5	60.3	86.8	57.3	81.3	100.0%
Retain 640 Tokens(↓ 77.8%)						
FastV(ECCV24)	58.9	58.1	79.5	53.5	77.0	94.1%
PruMerge+(2024.05)	60.8	54.9	85.3	57.3	78.2	96.6%
TRIM(COLING25)	62.1	54.8	86.9	55.8	78.3	97.0%
DART(EMNLP25)	61.3	59.5	85.0	57.1	78.3	98.1%
SparseVLM(ICML25)	61.2	59.7	85.3	58.6	79.2	99.0%
VisionZip(CVPR25)	61.2	59.9	86.0	58.1	79.1	99.0%
CDPruner(NeurIPS25)	62.6	58.4	87.3	57.5	79.9	99.2%
CrisPrune	62.7	60.1	87.2	58.3	80.1	**100.1%**
Retain 320 Tokens(↓ 88.9%)						
FastV(ECCV24)	49.8	52.2	49.5	42.5	61.5	74.6%
TRIM(COLING25)	59.9	50.2	86.5	51.0	74.9	92.6%
PruMerge+(2024.05)	58.8	54.0	79.5	55.6	75.3	92.8%
SparseVLM(ICML25)	57.9	56.5	76.9	56.7	74.6	93.1%
DART(EMNLP25)	59.5	57.6	81.0	55.7	75.7	94.9%
VisionZip(CVPR25)	58.9	58.8	82.3	55.6	76.2	95.5%
CDPruner(NeurIPS25)	61.6	57.4	87.2	55.7	78.4	97.6%
CrisPrune	61.5	58.6	87.3	57.2	78.6	**98.5%**
Retain 160 Tokens(↓ 94.4%)						
PruMerge+(2024.05)	56.2	50.3	71.1	48.9	70.5	85.3%
TRIM(COLING25)	57.4	45.8	84.8	45.2	71.0	87.4%
VisionZip(CVPR25)	55.2	55.0	74.9	50.4	71.4	88.3%
DART(EMNLP25)	56.8	54.9	75.3	53.6	72.5	90.3%
CDPruner(NeurIPS25)	60.8	55.4	86.8	53.8	76.7	95.5%
CrisPrune	61.1	57.7	86.7	56.3	77.2	**97.3%**

Table 2: **Performance comparison of different pruning methods on LLaVA-NeXT-7B.** The vanilla number of visual tokens is 2880. Rel represents the average percentage of performance maintained at the corresponding token retention level.

Method	GQA	MME	POPE	MMB-CN	MMB	Rel
Vanilla, Retain 100% Tokens						
Qwen2.5-VL-7B	60.5	2331	86.2	80.1	83.2	100.0%
Retain 33.3% Tokens						
HiPrune(AAAI26)	58.9	2303	85.1	79.5	82.6	98.7%
CrisPrune	59.5	2293	85.4	80.5	82.1	**99.0%**
Retain 22.2% Tokens						
HiPrune(AAAI26)	57.1	2189	84.0	77.6	80.3	95.8%
CrisPrune	57.8	2243	84.0	78.4	80.9	**96.9%**
Retain 11.1% Tokens						
HiPrune(AAAI26)	52.5	2000	79.8	73.7	76.1	89.7%
CrisPrune	54.5	2065	80.7	74.6	76.9	**91.6%**

Table 3: **Performance comparisons on Qwen2.5-VL-7B-Instruct.** Rel represents the average percentage of performance maintained at the corresponding token retention level.

that maximizes this objective is NP-hard, we employ the fast greedy algorithm based on Cholesky factorization (Chen et al., 2018). This approach iteratively selects tokens that maximize the marginal gain in log-determinant, ensuring high efficiency during inference.

4 Experiment

This section presents a comprehensive evaluation of CrisPrune across a wide range of MLLMs and diverse benchmarks. We benchmark our method against leading state-of-the-art methods, conduct rigorous ablation studies to justify our design decisions, and offer a detailed examination of computational efficiency.

4.1 Results on Image Understanding

Datasets and model architectures. We conduct our evaluation on 9 widely-used benchmarks, encompassing conventional Visual Question Answering (VQA) datasets such as VQAv2(Goyal et al., 2017), GQA(Hudson and Manning, 2019), ScienceQA(Lu et al., 2022), and TextVQA(Singh et al., 2019), as well as other prominent multimodal benchmarks including POPE(Li et al., 2023b), MME(Chaoyou et al., 2023), MMBench(Liu et al., 2024c), MMBench-CN(Liu et al., 2024c), and MMVet(Yu et al., 2023). We apply our method to a variety of MLLM architectures, including LLaVA-1.5(Liu et al., 2024a), LLaVA-NeXT(Liu et al., 2024b) and other prevalent models such as Qwen2.5-VL(Bai et al., 2025). Further details are provided in the Appendix B.

Results on LLaVA-1.5. As presented in Table 1, we apply our proposed method to LLaVA-1.5 to demonstrate its performance on image understanding tasks. For a comprehensive comparison, we report performance as a relative percentage, where the accuracy of the original, unpruned model serves as the 100% upper bound. Following the prior work(Zhang et al., 2025a), we evaluate the efficacy of CrisPrune under three visual token budgets (128, 64, and 32).

When the number of visual tokens is reduced from 576 to 128, CrisPrune maintains 99.4% of the original average accuracy without requiring any additional training, outperforming VisionZip and DivPrune by 1.8% and 1.9%. When the number of visual tokens further decreases to 64, CrisPrune only decreases the original performance by 2.4%. Moreover, when pruned aggressively to only 32 tokens, our method surpasses VisionZip and CDPruner by even more significant margins of 7.1% and 2.5% in average accuracy.

Results on LLaVA-NeXT. While utilizing higher-resolution inputs significantly enhances MLLM performance across various tasks, it inevitably leads to a substantial increase in visual tokens, thereby exacerbating computational overhead. To

Method	TGIF-QA		MSVD-QA		MSRVTT-QA		Average	
	Acc.	Score	Acc.	Score	Acc.	Score	Acc.	Score
*Vanilla, 2048 Tokens (**100%**)*								
Video-LLaVA	18.9	2.53	71.7	3.95	57.5	3.50	49.4	3.33
Retain 455 Tokens(↓ 77.8%)								
FastV(ECCV24)	19.2	2.50	69.1	3.91	54.4	3.42	47.6	3.28
VisPruner(CVPR25)	18.0	2.49	70.2	3.95	56.7	3.50	48.3	3.31
CrisPrune	16.9	2.46	71.1	3.98	57.2	3.51	**48.4**	**3.32**
Retain 227 Tokens(↓ 88.9%)								
FastV(ECCV24)	14.3	2.42	68.9	3.90	53.0	3.40	45.4	3.24
VisPruner(CVPR25)	15.9	2.41	69.3	3.92	55.6	3.45	46.9	3.26
CrisPrune	15.5	2.43	70.4	3.93	56.3	3.45	**47.4**	**3.27**
Retain 114 Tokens(↓ 94.4%)								
FastV(ECCV24)	10.6	2.29	64.1	3.78	52.4	3.39	42.4	3.15
VisPruner(CVPR25)	14.1	2.35	65.4	3.79	54.1	3.41	44.5	3.18
CrisPrune	14.7	2.40	69.8	3.91	55.0	3.42	**46.5**	**3.24**

Table 4: **Performance comparison of different pruning methods on Video-LLaVA-7B.** Performance was assessed using the initial 1,000 samples from each benchmark, with gpt-3.5-turbo employed for scoring.

Method	Token	Storage (MB)	GPU Memory (GB)	CUDA Time (ms)
LLaVA-NeXT-7B	2880	1440	17.0	313
FastV(ECCV24)	640	380	16.9	148
CrisPrune	640	**351**	**15.6**	**145**

Table 5: **Resource efficiency analysis with LLaVA-NeXT-7B on POPE.** We report the number of visual tokens, storage size, peak GPU memory usage, and single-pass CUDA time.

Method	Token	FLOPS ↓	Total Time ↓	Δ	Prefilling Time ↓	Δ
LLaVA-NeXT-7B	2880	43.6T	2293s	—	218ms	—
FastV(ECCV24)	160	6.3T	1792s	1.3×	119ms	1.8×
CrisPrune	160	3.3T	**933s**	**2.5**×	**20.0ms**	**10.9**×

Table 6: **Comparison of inference speed and computational cost on LLaVA-NeXT-7B.** We report total FLOPS, total generation time, and prefilling time. The speed-up ratio (Δ) is calculated against the baseline.

validate the efficacy of CrisPrune in this context, we apply it to LLaVA-NeXT, a model designed for high-resolution inputs capable of processing up to 2,880 visual tokens. We evaluate our method under three token budgets (640, 320, and 160).

As detailed in Table 2, CrisPrune demonstrates superior performance across all three settings. Even with 77.8% of tokens pruned, CrisPrune maintains performance comparable to, or even slightly exceeding, that of the original LLaVA-NeXT. Notably, with only 320 tokens retained, our method preserves 98.5% of the original average accuracy, outperforming VisionZip by 3.0%. Furthermore, even under the aggressive pruning setting of 160 tokens, CrisPrune sustains a robust 97.3% average accuracy, surpassing VisionZip and CDPruner by significant margins of 9.0% and 1.8%, respectively. These results underscore the robustness and effectiveness of CrisPrune in high-resolution scenarios. The results for LLaVA models at different scales can be found in the Appendix C.1.

Results on Qwen2.5-VL. To validate the versatility of CrisPrune, we integrated it into the advanced open-source Qwen2.5-VL architecture. This is a crucial test, as many competing approaches that rely on specific components like a CLIP(Radford et al., 2021) text encoder or special tokens are often incompatible with such models. Since Qwen2.5-VL generates variable token lengths due to dynamic resolution, we follow the protocol of HiPrune (Liu et al., 2025) and adopt percentage-based retention ratios (i.e., 33.3%, 22.2%, and 11.1%).

As illustrated in Table 3, CrisPrune achieves superior performance over HiPrune across all tested pruning ratios. Our method maintains 99.0% of the vanilla model's capabilities when pruning 66.7% of the tokens. Furthermore, at the extreme pruning ratio where only 11.1% of tokens are kept, CrisPrune still achieves a 91.6% relative score, demonstrating its strong adaptability and effectiveness on dynamic resolution models like Qwen2.5-VL. The results for Qwen2.5-VL-3B-Instruct can be found in the Appendix C.2.

4.2 Results on Video Understanding

Datasets and Model architectures. We evaluate our proposed method on three widely-used video question answering benchmarks: TGIF-QA(Jang et al., 2017), MSVD-QA(Xu et al., 2017), and MSRVTT-QA(Xu et al., 2017). Our approach is applied to the Video-LLaVA(Lin et al., 2023) model. Following prior works(Lin et al., 2023; Maaz et al., 2023), we employ a GPT-assisted evaluation protocol where scores assigned by ChatGPT serve as the primary performance metric.

Results on Video-LLaVA. Due to the commercial API usage limits, we follow(Chen et al., 2024a; Zhang et al., 2025a) to use the first 1K samples from each benchmark in our experiments. As shown in Table 4, at a moderate compression level (retaining 455 tokens), our method achieves an average accuracy of 48.4% and a GPT score of 3.32, effectively retaining 98.0% of the original performance while reducing the visual context by nearly

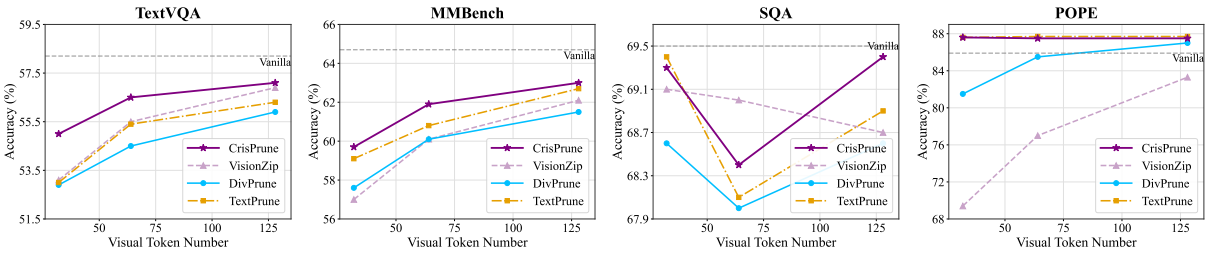

Figure 5: **Ablation study of CrisPrune components.** We evaluate the performance of different pruning strategies on LLaVA-1.5-7B. TextPrune refers to a variant that relies solely on dual-source text relevance; VisionZip represents pruning based exclusively on visual attention; and DivPrune utilizes Max-Min distance for diversity selection.

80%. The superiority of CrisPrune becomes most pronounced under extreme compression (retaining only 114 tokens). While competing methods suffer significant degradation, with FastV dropping to 85.8% and VisPruner to 90.1%, CrisPrune sustains a remarkable 94.1% relative performance.

4.3 Efficiency Analysis

To assess computational efficiency of our proposed method, we analyze metrics including total inference time, pre-filling time, FLOPS, cache storage, GPU memory, and CUDA latency. Experiments were conducted on LLaVA-NeXT-7B against FastV (Chen et al., 2024a) using the POPE on a single NVIDIA A100 GPU.

As detailed in Table 5, retaining 640 tokens yields a 2.2× reduction in CUDA latency with lower cache and GPU memory consumption compared to FastV. Furthermore, Table 6 highlights substantial efficiency gains: our method reduces FLOPS by a factor of 13, lowers total inference time by 2.5×, and reduces pre-filling time by 10.9×. Crucially, CrisPrune performs pruning entirely before the LLM, enabling full compatibility with optimized attention mechanisms like FlashAttention(Dao et al., 2022). This offers a distinct advantage over methods dependent on internal text-vision attention scores, which are often incompatible with such frameworks.

4.4 Ablation Study

We further conduct an ablation study to investigate the contribution of each component in CrisPrune, as illustrated in Figure 5. We evaluate the performance on LLaVA-1.5-7B across four benchmarks under varying visual token budgets. Here, TextPrune serves as a visually-degraded variant of CrisPrune, which determines token quality solely based on dual-source text relevance. VisionZip(Yang et al., 2025) represents a vision-only baseline relying on attention scores, and DivPrune(Alvar et al., 2025) focuses purely on diversity via Max-Min Diversity Problem (MMDP)(Resende et al., 2010).

The results reveal several critical insights. First, DivPrune consistently yields the lowest performance, indicating that pursuing diversity without a robust quality metric introduces noise and degrades reasoning. Second, TextPrune generally outperforms VisionZip, particularly on text-oriented tasks like TextVQA. This suggests that our dual-source text relevance provides more precise guidance for token selection than query-agnostic visual attention. Most importantly, CrisPrune achieves the best performance across the majority of settings. By reintegrating *Intrinsic Visual Saliency* with text relevance, CrisPrune recovers structural details that might be overlooked by text alignment alone. This validates the effectiveness of our joint modeling approach, where visual saliency and text relevance complement each other to identify the most informative subset via DPP. More ablation studies on hyperparameter sensitivity and prompts setting can be found in the Appendix C.3.

5 Conclusion

In this paper, we introduced CrisPrune, a plug-and-play visual token pruning framework. By integrating intrinsic visual saliency with dual-source text relevance, our approach robustly preserves holistic visual integrity while mitigating attention sinks and preserving essential background context. Furthermore, we reformulated the selection process using a determinantal point process to maximize conditional diversity, effectively eliminating spatial redundancy. Extensive experiments demonstrate that CrisPrune significantly outperforms state-of-the-art methods. We believe this work provides a robust and efficient solution for deploying high-resolution MLLMs in resource-constrained environments.

6 Limitations

While CrisPrune demonstrates superior performance and robustness, several limitations remain. First, although we employ a fast greedy algorithm, the DPP still introduces a computational overhead for kernel construction and selection. Second, while our general semantic priors prove effective across diverse tasks, the current prompt set is manually curated. Exploring automated prompt generation or learning task-specific priors could further enhance adaptability. Finally, the introduction of hyperparameters λ and α requires careful tuning to balance visual and textual guidance. Adaptive mechanisms to dynamically adjust these parameters would be a valuable direction for future work.

7 Ethical Considerations

We strictly adhere to the ethical guidelines of the research community. In this work, all experiments are conducted exclusively using publicly available datasets and open-source models, ensuring the transparency and reproducibility of our findings. We have not collected any new data involving human subjects or personally identifiable information. Furthermore, our proposed method, CrisPrune, focuses on the efficiency of visual token processing and does not modify the underlying safety alignment of the original MLLMs.

8 Acknowledgments

This work was supported in part by National Natural Science Foundation of China under Grant U23A20382, 62501422; in part by Shanghai Pujiang Program under Grant 25PJD128; in part by the Fundamental Research Funds for the Central Universities, and in part by Ant Group Research Intern Program.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

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

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, and 1 others. 2023a. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

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

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*.

Fu Chaoyou, Chen Peixian, Shen Yunhang, Qin Yulei, Zhang Mengdan, Lin Xu, Yang Jinrui, Zheng Xiawu, Li Ke, Sun Xing, and 1 others. 2023. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 3.

David Chen and William B Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th annual meeting of the association for computational linguistics: human language technologies*, pages 190–200.

Laming Chen, Guoxin Zhang, and Eric Zhou. 2018. Fast greedy map inference for determinantal point process to improve recommendation diversity. *Advances in neural information processing systems*, 31.

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.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, and 1 others. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24185–24198.

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. 2022. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in neural information processing systems*, 35:16344–16359.

Alexey Dosovitskiy. 2020. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*.

Yuxiang Duan, Ao Li, Yingqin Li, Luyu Li, and Pengwei Wang. 2025. Gridprune: From" where to look" to" what to select" in visual token pruning for mllms. *arXiv preprint arXiv:2511.10081*.

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

Jiayi Han, Liang Du, Yiwen Wu, Xiangguo Zhou, Hongwei Du, and Weibo Zheng. 2025. Adafv: Rethinking of visual-language alignment for vlm acceleration. *arXiv preprint arXiv:2501.09532*.

Grant Van Horn, Oisin Mac Aodha, Yang Song, Yin Cui, Chen Sun, Alex Shepard, Hartwig Adam, Pietro Perona, and Serge Belongie. 2018. The iNaturalist species classification and detection dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 8769–8778.

Drew A Hudson and Christopher D Manning. 2019. Gqa: A new dataset for real-world visual reasoning and compositional question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6700–6709.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Ivan Krasin, Tom Duerig, Neil Alldrin, Vittorio Ferrari, Sami Abu-El-Haija, Alina Kuznetsova, Hassan Rom, Jasper Uijlings, Stefan Popov, Andreas Veit, and 1 others. 2017. Openimages: A public dataset for large-scale multi-label and multi-class image classification. *Dataset available from https://github. com/openimages*, 2(3):18.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023a. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR.

Yanshu Li, Jianjiang Yang, Zhennan Shen, Ligong Han, Haoyan Xu, and Ruixiang Tang. 2025. Catp: Contextually adaptive token pruning for efficient and enhanced multimodal in-context learning. *arXiv preprint arXiv:2508.07871*.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer.

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*.

Yuncheng Li, Yale Song, Liangliang Cao, Joel Tetreault, Larry Goldberg, Alejandro Jaimes, and Jiebo Luo. 2016. Tgif: A new dataset and benchmark on animated gif description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4641–4650.

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*.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*, pages 740–755. Springer.

Zhihang Lin, Mingbao Lin, Luxi Lin, and Rongrong Ji. 2025. 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.

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

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

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

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. 2023. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*.

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

Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, 42(3):145–175.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, and 1 others. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR.

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

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. 2025. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22857–22867.

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.

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.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, and 1 others. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

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

Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, pages 1645–1653.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

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 Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan Wang. 2023. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986.

Qizhe Zhang, Aosong Cheng, Ming Lu, Renrui Zhang, Zhiyong Zhuo, Jiajun Cao, Shaobo Guo, Qi She, and Shanghang Zhang. 2025a. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20857–20867.

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

Shaolei Zhang, Qingkai Fang, Zhe Yang, and Yang Feng. 2025c. Llava-mini: Efficient image and video large multimodal models with one vision token. *arXiv preprint arXiv:2501.03895*.

Yuan Zhang, Chun-Kai Fan, Junpeng Ma, Wenzhao Zheng, Tao Huang, Kuan Cheng, Denis A Gudovskiy, Tomoyuki Okuno, Yohei Nakata, Kurt Keutzer, and 1 others. 2024. Sparsevlm: Visual token sparsification for efficient vision-language model inference. In *Forty-second International Conference on Machine Learning*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Preprint*, arXiv:2306.05685.

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, 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*.

Algorithm 1 Pseudocode for CrisPrune

```
# Inputs: Visual tokens X, Attention A
      Instruction T, General Prompts P

# Robust Min-Max Normalization
function Norm(v):
    return (v - v.min()) / (v.max() -
        v.min() + ϵ)

# 1. Robust Intrinsic Visual Saliency
attn = A.mean(dim=0)
v_min, v_rob = attn.min(), Quantile(
    attn, 0.98)
s_vis = Clamp((attn - v_min) / (v_rob
    - v_min), 0, 1)

# 2. Dual-Source Semantic Relevance
r_inst = Norm(CosineSim(X, T))
r_gen = Norm(CosineSim(X, Encode(P)).
    max(dim=-1))
s_sem = Max(r_inst, λ * r_gen)

# 3. Quality Kernel Construction
q = α * s_sem + (1 - α) * s_vis
Sim = MatMul(X, X.T)
# Decompose kernel: L_ij = q_i *
    Sim_ij * q_j
L = Outer(q, q) * Sim

# 4. Fast Greedy MAP Inference
S = []
d_sq = L.diagonal().clone()
for i = 1 to K:
    j = argmax(d_sq)
    S.append(j)
    UpdateCholeskyFactors(L, S, d_sq)
    d_sq[j] = -∞

return X[S]
```

A Pseudo-code for CrisPrune

In Algorithm 1, we provide a pseudo-code for CrisPrune written in PyTorch style to better explain our method. This example is adapted from LLaVA series.

B Experiment Detail

B.1 Model Architectures

We apply our method to a variety of MLLM architectures, including LLaVA-1.5(Liu et al., 2024a), LLaVA-NeXT(Liu et al., 2024b), Video-LLaVA(Lin et al., 2023), and Qwen2.5-VL(Bai et al., 2025).

LLaVA-1.5. LLaVA stands as a foundational paradigm for open-source vision-language models, known for its efficient architecture and robust performance. It integrates a pre-trained CLIP(Radford et al., 2021) vision encoder with a Vicuna(Zheng et al., 2023) LLM to handle multimodal generation. A key upgrade of LLaVA-1.5 from the original LLaVA is the replacement of the linear projection layer with a two-layer MLP, which, combined with increased input resolution and a broadened instruction-tuning dataset, yields significant performance gains.

LLaVA-NeXT. Building upon LLaVA-1.5, LLaVA-NeXT (also known as LLaVA-1.6) introduces critical architectural enhancements to overcome resolution bottlenecks. Its primary advancement is the support for dynamic high-resolution imagery, enabling the processing of visual inputs at multiple scales—up to 4x the resolution of its predecessor. This "AnyRes" capability, coupled with a high-quality visual instruction dataset, results in marked improvements in fine-grained recognition and logical reasoning.

Video-LLaVA. Video-LLaVA challenges the conventional separation of image and video encoders by proposing a unified visual representation. Its core contribution lies in unifying the tokenization process, projecting visual inputs into a shared semantic space compatible with the LLM. By training on a combined corpus of images and videos, Video-LLaVA aligns feature spaces effectively, fostering a synergistic learning process that enhances performance across both modalities.

Qwen2.5-VL. Qwen2.5-VL is a state-of-the-art model featuring a native dynamic-resolution Vision Transformer (ViT)(Dosovitskiy, 2020) and absolute time embeddings. It adopts a modernized ViT architecture with window attention, SwiGLU activation, and RMSNorm, aligned with the Qwen2.5 LLM. This design allows for the processing of variable-sized images and long-duration videos at their native scale. Consequently, Qwen2.5-VL demonstrates exceptional capabilities in fine-grained visual grounding, document parsing, and long-context video comprehension.

B.2 Comparison methods

We benchmark our method against a comprehensive set of state-of-the-art baselines, including FastV(Chen et al., 2024a), SparseVLM(Zhang et al., 2024), TRIM(Song et al., 2025), PruMerge+(Shang et al., 2025), DART(Wen et al., 2025), VisionZip(Yang et al., 2025), DivPrune(Alvar et al., 2025), GridPrune(Duan et al., 2025), CDPruner(Zhang et al., 2025b), VisPruner(Zhang et al., 2025a), and HiPrune(Liu et al., 2025).

Method	MME	TextVQA	SQA	MMB-CN	GQA	MMBench	POPE	VQA^{v2}	MM-Vet	Avg. Rel (%)
*Vanilla, 576 Tokens (**100%**)*										
LLaVA-1.5-13B	1531.2	61.2	72.8	63.5	63.3	68.5	86.0	80.0	36.2	100.0%
	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	100.0%	
LLaVA-1.5-7B	1506.5	58.2	69.5	58.1	61.9	64.7	85.9	78.5	31.3	95.3%
	98.4%	95.1%	95.5%	91.5%	97.8%	94.5%	99.9%	98.1%	86.5%	
Retain 128 Tokens (↓ 77.8%)										
CrisPrune	1475.4	58.8	73.0	61.8	59.8	68.0	86.9	78.0	37.7	**98.5%**
	96.4%	96.1%	100.3%	97.3%	94.5%	99.3%	101.0%	97.5%	104.1%	
Retain 64 Tokens (↓ 88.9%)										
CrisPrune	1477.2	58.3	72.5	61.5	59.1	66.6	86.4	77.2	38.9	**98.2%**
	96.5%	95.3%	99.6%	96.9%	93.4%	97.2%	100.5%	96.5%	107.5%	
Retain 32 Tokens (↓ 94.4%)										
CrisPrune	1420.1	57.2	72.6	60.0	58.8	65.1	85.4	75.7	34.6	**95.3%**
	92.7%	93.5%	99.7%	94.5%	92.9%	95.0%	99.3%	94.6%	95.6%	

Table 7: **Performance of CrisPrune on LLaVA-1.5-13B.** The vanilla number of visual tokens is 576. The first line of each method shows the raw benchmark accuracy, and the second line is the proportion relative to the upper limit. The final column shows the average performance relative to the original model.

B.3 Evaluation Benchmark

We conduct experiments on a total of 12 widely used visual understanding benchmarks, including 9 image benchmarks and 3 video benchmarks.

B.3.1 Image Benchmark

We conducted experiments on 9 widely used visual understanding benchmarks. We use the official implementation of inference settings and evaluation metrics in LLaVA-1.5.

MME. The MME benchmark(Chaoyou et al., 2023) provides a comprehensive evaluation of multimodal models by assessing both perceptual and cognitive capabilities across 14 distinct subtasks. Designed as binary judgment problems, these subtasks are organized into two primary categories: perception, which includes Optical Character Recognition (OCR), coarse-grained recognition, and fine-grained recognition; and cognition, which encompasses commonsense reasoning, numerical calculation, text translation, and code reasoning.

TextVQA. The TextVQA(Singh et al., 2019) benchmark evaluates a model's ability to read and reason about text embedded in images. Sourced from Open Images v3(Krasin et al., 2017), the dataset features text-heavy scenes such as signage and packaging. It requires models to perform OCR and integrate this with visual context to answer questions. We report results on the validation split.

ScienceQA. Designed to assess zero-shot generalization, ScienceQA(Lu et al., 2022) covers diverse scientific domains through a multiple-choice format. The benchmark is structured hierarchically, spanning 26 topics, 127 categories, and 379 distinct skills. For our evaluation, we utilize the test split containing image contexts.

MM-Vet. MM-Vet(Yu et al., 2023) evaluates the integration of six core vision-language capabilities: recognition, OCR, knowledge, language generation, spatial awareness, and mathematics. Comprising 218 image-question pairs, it systematically combines these skills into 16 composite tasks. A distinguishing feature is its use of an LLM-based evaluator (ChatGPT) to provide a unified metric for open-ended responses.

MMBench. MMBench(Liu et al., 2024c) offers a robust evaluation pipeline through a three-tier taxonomy, rooted in foundational perception and reasoning abilities. It assesses models using a circular evaluation strategy to ensure robustness. We evaluate on both the English version (4,377 pairs) and the Chinese version (**MMBench-CN**, 4,329 pairs).

GQA. Focusing on compositional reasoning, GQA(Hudson and Manning, 2019) leverages scene graphs from the Visual Genome dataset(Krishna et al., 2017) to provide rigorous assessment. It requires models to understand objects, attributes, and spatial relationships to answer questions. We utilize the test-dev balanced split for evaluation.

Method	MME	TextVQA	SQA	MMB-CN	GQA	MMBench	POPE	VQA^{v2}	MM-Vet	Avg. Rel (%)
*Vanilla, 2880 Tokens (**100%**)*										
LLaVA-NeXT-13B	1580.1 100.0%	63.2 100.0%	73.1 100.0%	61.2 100.0%	64.4 100.0%	68.5 100.0%	85.3 100.0%	82.3 100.0%	45.0 100.0%	100.0%
Retain 640 Tokens (↓ 77.8%)										
CrisPrune	1579.8 100.0%	62.3 98.6%	71.5 97.8%	63.1 103.1%	64.0 99.4%	68.8 100.4%	87.3 102.3%	81.2 98.7%	44.5 98.9%	**99.9%**
Retain 320 Tokens (↓ 88.9%)										
CrisPrune	1523.3 96.4%	60.7 96.0%	70.7 96.7%	62.8 102.6%	63.3 98.3%	67.7 98.8%	87.2 102.2%	79.8 97.0%	42.9 95.3%	**98.1%**
Retain 160 Tokens (↓ 94.4%)										
CrisPrune	1493.2 94.5%	59.3 93.8%	71.2 97.4%	61.8 101.0%	62.4 96.9%	66.8 97.5%	87.9 103.0%	78.0 94.8%	41.3 91.8%	**96.7%**

Table 8: **Performance of CrisPrune on LLaVA-NeXT-13B.** The vanilla number of visual tokens is 2880. The first line of each method shows the raw benchmark accuracy, and the second line is the proportion relative to the upper limit. The final column shows the average performance relative to the original model.

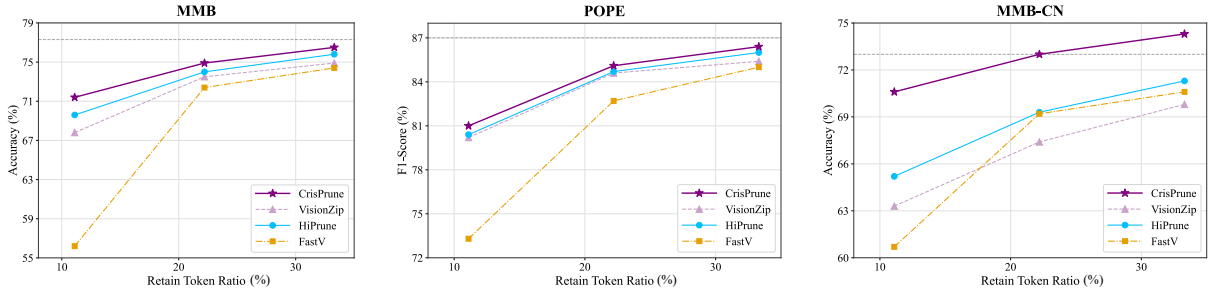

Figure 6: **Performance of CrisPrune on Qwen2.5-VL-3B-Instruct across different token retention ratios.** The grey dashed line represents the performance of the vanilla model.

VQAv2. VQAv2(Goyal et al., 2017) is a standard benchmark for open-ended visual question answering, built upon the MSCOCO dataset(Lin et al., 2014). It addresses the language bias issues of its predecessor by employing a balanced pair design, where each question is associated with complementary images that yield different answers. We use the test-dev split.

POPE. POPE(Li et al., 2023b) is specifically designed to quantify object-level hallucination in Large Vision-Language Models (LVLMs). It formulates evaluation as a binary classification task, probing the model's ability to correctly identify the presence or absence of objects in MSCOCO(Lin et al., 2014) images.

B.3.2 Video Benchmark

Our approach is implemented on Video-LLaVA and evaluated on three major VideoQA benchmarks. Following prior work(Lin et al., 2023; Maaz et al., 2023), we use GPT-3.5-Turbo for automated evaluation. In line with (Chen et al., 2024a; Zhang et al., 2025a) and due to API usage constraints, we conduct experiments on the initial 1,000 samples of each benchmark.

TGIF-QA. TGIF-QA(Jang et al., 2017) extends static image-based VQA to the temporal domain, evaluating a model's ability to perform spatio-temporal reasoning. This task necessitates understanding both spatial details within individual frames and temporal dynamics across them. Constructed from 72,000 animated GIFs from the TGIF dataset(Li et al., 2016), the benchmark comprises 165,000 question-answer pairs.

MSVD-QA. Adapted from the MSVD video captioning dataset(Chen and Dolan, 2011), MSVD-QA(Xu et al., 2017) serves as a standard benchmark for assessing video comprehension. Its question-answer pairs are automatically generated from ground-truth video descriptions to probe the understanding of video content. In total, the benchmark consists of 1,970 short video clips and approxi-

Prompt Format	MME	TextVQA	POPE	Avg. Rel
Baseline (LLaVA-1.5-7B, 576 Tokens)				
Vanilla	1506.5	58.2	85.9	100.0%
Retain 128 Tokens (↓ 77.8%)				
Full Sentence	1454.7	57.1	87.5	98.8%
Concise Phrase	1448.9	57.1	87.3	98.6%
Retain 64 Tokens (↓ 88.9%)				
Full Sentence	1423.8	56.5	87.5	97.8%
Concise Phrase	1425.7	56.6	87.6	98.0%
Retain 32 Tokens (↓94.4%)				
Full Sentence	1398.5	55.0	87.8	96.5%
Concise Phrase	1378.7	55.0	87.4	95.9%

Table 9: **Ablation study on prompt granularity.** We compare the effectiveness of using full sentences (e.g., "*All objects in the image*") versus concise phrases (e.g., "*All objects*") as general semantic priors. Experiments are conducted on LLaVA-1.5-7B.

Prompt Count	MME	TextVQA	POPE	Avg. Rel
Baseline (LLaVA-1.5-7B, 576 Tokens)				
Vanilla	1506.5	58.2	85.9	100.0%
Retain 128 Tokens (↓ 77.8%)				
$K = 3$	1454.7	57.1	87.5	98.8%
$K = 5$	1453.9	57.1	87.5	98.8%
$K = 10$	1447.1	57.0	87.4	98.6%
Retain 64 Tokens (↓ 88.9%)				
$K = 3$	1423.8	56.5	87.5	97.8%
$K = 5$	1424.5	56.6	87.5	97.9%
$K = 10$	1425.0	56.5	87.7	97.9%
Retain 32 Tokens (↓ 94.4%)				
$K = 3$	1398.5	55.0	87.8	96.5%
$K = 5$	1398.4	55.1	87.5	96.5%
$K = 10$	1395.2	54.7	87.2	96.0%

Table 10: **Ablation study on the number of general semantic prompts (K).** We compare our default setting ($K = 3$) with expanded sets ($K = 5, 10$) that include additional background and fine-grained descriptions. Experiments are conducted on LLaVA-1.5-7B.

mately 50,500 question-answer pairs.

MSRVTT-QA. Built upon the large-scale MSRVTT dataset(Xu et al., 2016), MSRVTT-QA(Xu et al., 2017) is a comprehensive dataset for video question answering. It encompasses 10,000 video clips and nearly 243,000 question-answer pairs. The questions cover diverse categories (e.g., *what*, *who*, *how*), requiring models to effectively process and integrate visual and temporal information.

B.4 Implementation Details

For image and video benchmarks, we utilize the official implementations of LLaVA and Video-LLaVA, respectively. For advanced architectures such as Qwen2.5-VL, we adopt the settings from HiPrune(Liu et al., 2025). Inference was conducted on NVIDIA A100 GPUs; specifically, we employed 8 GPUs for GQA and VQAv2 benchmarks, and a single GPU for all other datasets.

C Additional Experiments

C.1 Performance on LLaVA at Different Scales

Results on LLaVA-1.5-13B. In the main paper, we demonstrate the effectiveness of our model on the 7B scale in Table 1; this section extends our analysis to the 13B model. As shown in Table 7, we evaluate our method across three vision token counts (128, 64, and 32). The results indicate that even with a limited budget of 32 visual tokens, our method preserves 95.3% of the baseline performance. At 64 tokens, the model maintains an impressive 98.2% performance. Notably, when retaining 64 or 128 tokens, the compressed 13B model outperforms the standard 7B model.

Results on LLaVA-NeXT-13B. We demonstrate the effectiveness of CrisPrune on the LLaVA-NeXT-7B model across a comprehensive suite of benchmarks in Table 2. Our evaluation is further extended to the more powerful and challenging LLaVA-1.6-13B baseline to demonstrate the generalizability of our method. As presented in Table 8, our approach shows remarkable performance preservation across different compression levels. At a modest 640 tokens, the model retains 99.9% of its original performance. This robustness is maintained at 320 tokens, achieving 98.1% of the baseline. Even under aggressive compression to just 160 tokens, the model still delivers an impressive 96.7% of its full capability, underscoring its high efficiency and robustness.

C.2 Performance on Qwen-VL at Different Scales

Results on Qwen2.5-VL-3B-Instruct. In the main paper, we demonstrate the effectiveness of our model on the 7B scale in Table 3; this section extends our analysis to the 3B model. Figure 6 reveals that the superiority of CrisPrune becomes increasingly pronounced as visual tokens are reduced across all benchmarks. On MMBench-CN, specifically under aggressive compression, our method maintains robust performance while competing ap-

Balance Factor (α)	MME	TextVQA	POPE	MMBench	SQA	MMB-CN	MM-Vet	Avg. Rel
Retain 128 Tokens(↓ 77.8%)								
$\alpha = 0.1$	1446.0	57.3	86.4	62.5	68.8	56.7	33.3	99.2%
$\alpha = 0.3$	1443.5	57.2	87.3	62.9	68.5	56.9	35.5	100.4%
$\alpha = 0.5$	1454.7	57.0	87.3	63.0	69.4	56.4	34.0	99.9%
$\alpha = 0.7$	1451.0	56.8	87.3	62.7	69.0	56.5	33.3	99.3%
$\alpha = 0.9$	1432.0	56.5	87.5	63.1	68.6	55.3	31.9	98.2%
Retain 64 Tokens(↓ 88.9%)								
$\alpha = 0.1$	1411.3	56.5	83.7	61.2	68.6	56.0	31.5	96.9%
$\alpha = 0.3$	1410.9	56.5	86.3	61.3	68.7	55.5	32.8	97.9%
$\alpha = 0.5$	1414.9	56.5	86.7	61.6	68.6	55.2	32.1	97.6%
$\alpha = 0.7$	1425.9	56.1	87.3	61.4	68.7	55.1	31.0	97.2%
$\alpha = 0.9$	1412.6	55.4	87.6	61.3	68.5	53.7	29.5	95.8%
Retain 32 Tokens(↓ 94.4%)								
$\alpha = 0.1$	1332.2	55.0	78.6	59.6	69.1	54.4	29.2	93.3%
$\alpha = 0.3$	1370.2	55.0	84.6	60.1	69.0	54.0	30.7	95.3%
$\alpha = 0.5$	1395.0	55.0	86.4	59.8	69.3	53.2	31.4	96.0%
$\alpha = 0.7$	1372.9	54.7	86.9	59.5	68.9	52.4	29.7	94.6%
$\alpha = 0.9$	1370.1	55.2	87.6	59.1	69.2	50.4	30.1	94.5%

Table 11: **Sensitivity analysis on the balance factor α.** This hyperparameter controls the trade-off between semantic relevance (text) and intrinsic visual saliency. A higher α indicates a stronger reliance on textual guidance. Experiments are conducted on LLaVA-1.5-7B.

proaches suffer severe degradation. These results validate the effectiveness and generalizability of CrisPrune on advanced architectures.

C.3 Additional Ablation Studies

Ablation study on prompt granularity. As shown in Table 9, employing Full Sentences yields marginally superior performance compared to Concise Phrases, though the difference is negligible. This indicates that our method is highly robust to prompt granularity; whether formatted as complete sentences or isolated keywords, the semantic embeddings successfully encapsulate the necessary background context. This validates that the efficacy of our approach stems from the intrinsic semantic information itself rather than overfitting to specific prompt templates. Furthermore, given that specific user instructions typically take the form of complete sentences, we adopt the full-sentence format for general priors to maintain consistency in the dual-source text relevance calculation.

Ablation study on the number of generic prompts. We investigate the impact of the prompt set size K on model performance, as summarized in Table 10. We expanded the default set ($K = 3$) to include additional descriptions such as background environments and colors ($K = 5, 10$). The results indicate that increasing K to 5 yields negligible gains ($\leq 0.1\%$), suggesting that our core prompts already capture the majority of essential visual information. However, expanding to $K = 10$ leads to a slight performance degradation, particularly under aggressive compression (e.g., Avg. Rel drops from 96.5% to 96.0% when retaining 32 tokens).

This validates that our selected three prompts cover the orthogonal semantic spaces required

Decay Factor	MME	TextVQA	POPE	Avg. Rel
Baseline (LLaVA-1.5-7B, 576 Tokens)				
Vanilla	1506.5	58.2	85.9	100.0%
Retain 128 Tokens (↓ 77.8%)				
$\lambda = 0.4$	1442.3	57.0	87.5	98.5%
$\lambda = 0.5$	1454.7	57.1	87.5	98.8%
$\lambda = 0.6$	1454.3	57.2	87.2	98.8%
Retain 64 Tokens (↓ 88.9%)				
$\lambda = 0.4$	1415.9	56.4	87.7	97.7%
$\lambda = 0.5$	1423.8	56.5	87.5	97.8%
$\lambda = 0.6$	1422.4	56.5	87.8	98.0%
Retain 32 Tokens (↓ 94.4%)				
$\lambda = 0.4$	1390.3	54.9	87.3	96.1%
$\lambda = 0.5$	1398.5	55.0	87.8	96.5%
$\lambda = 0.6$	1394.1	55.1	87.4	96.3%

Table 12: **Sensitivity analysis on the decay factor λ.** This parameter balances the influence of the general semantic prior relative to the specific instruction. Experiments are conducted on LLaVA-1.5-7B.

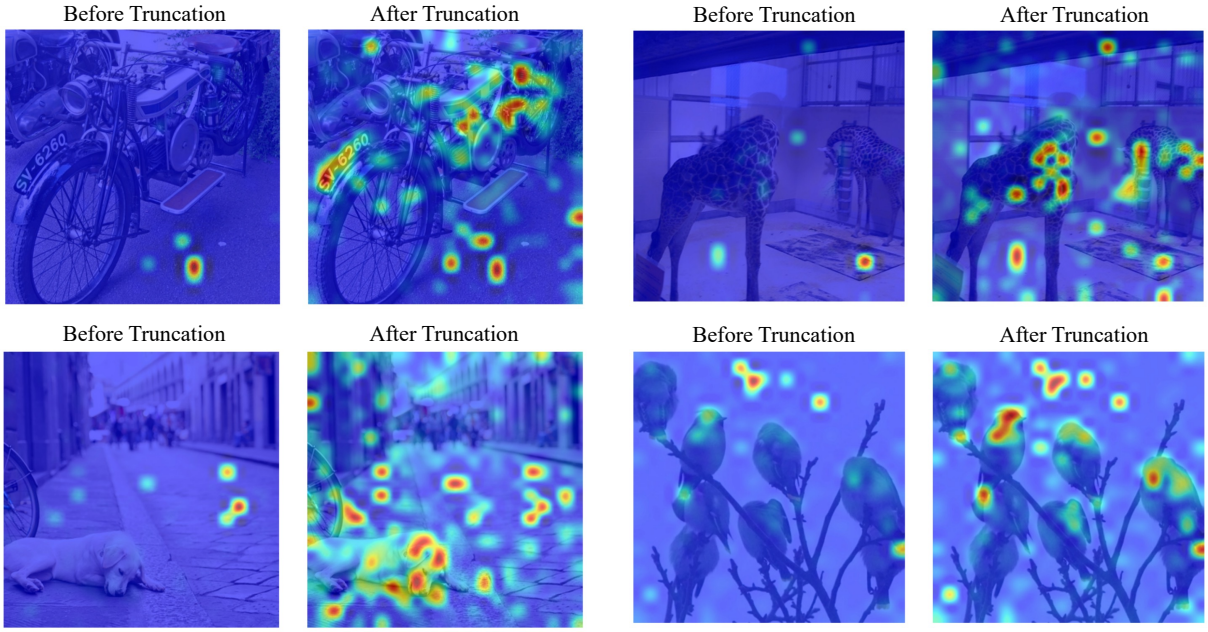

Figure 7: **Qualitative comparison of attention maps before and after Robust Truncation.** Each pair displays the raw attention map (Left, "Before Truncation") versus the map processed by our strategy (Right, "After Truncation").

for effective pruning. Specifically, these prompts encompass the dimensions of objects, text, and saliency. Excessive prompts introduce redundant or irrelevant semantic noise, diminishing the precision of the dual-source relevance score. Thus, we adopt $K = 3$ as the optimal trade-off between semantic coverage and computational efficiency.

Impact of Decay Factor λ. We analyze the sensitivity of the decay factor λ in Table 12, which controls the contribution of the general semantic prior. The results exhibit a clear trade-off. When λ is low, the model fails to fully leverage the background context provided by the general prior, leading to suboptimal performance (e.g., 96.1% at 32 tokens). Conversely, increasing λ to 0.6 yields slight gains at higher retention ratios but causes performance drops under aggressive compression (retaining 32 tokens). This suggests that an overly strong general prior may distract the model from the specific user instruction when the token budget is severely limited. Ultimately, $\lambda = 0.5$ achieves the most robust balance, delivering consistent gains across all settings and securing the highest performance in the most challenging scenario.

Impact of Balance Factor α. We investigate the effect of the balance factor α in Table 11, which governs the trade-off between dual-source semantic relevance and intrinsic visual saliency (see Eq. 6). The results reveal distinct performance patterns across tasks. Tasks dependent on specific object grounding, such as POPE, benefit from a higher α, as textual guidance helps pinpoint targets. Conversely, comprehensive reasoning tasks like MMBench-CN and MM-Vet suffer notable degradation when α is too high, indicating that over-reliance on text leads to the loss of critical background visual cues.

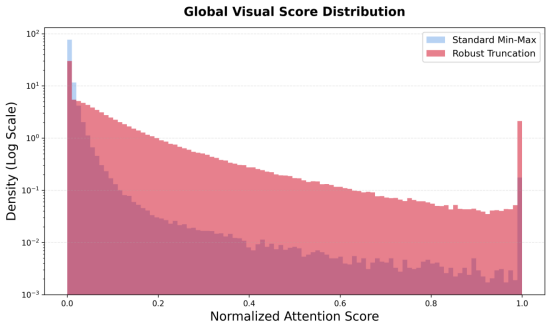

Figure 8: **Quantitative analysis of visual score distribution.** The histogram compares the probability density of attention scores under standard Min-Max normalization (blue) versus our Robust Truncation strategy (red).

Notably, at the most aggressive compression level (32 tokens), the extreme settings ($\alpha = 0.1$ or 0.9) perform poorly. In contrast, $\alpha = 0.5$ achieves the highest robustness (96.0%), validating that visual saliency and semantic relevance are equally indispensable for maintaining holistic model capabilities. Therefore, in practical applications, one may choose to tune this hyperparameter to prioritize either specific object grounding or broad scene

understanding based on specific needs.

D Additional Visualization

D.1 Visualization of Robust Normalization Strategy

To validate the effectiveness of our proposed *Robust Min-Max Normalization*, we provide both quantitative and qualitative analyses comparing it against the standard Min-Max approach. We first analyze the global impact on score distribution in Figure 8. The standard normalization (blue histogram) results in a heavy-tailed distribution where the vast majority of tokens are compressed into the near-zero range. This indicates that attention sinks stretch the scale, effectively "silencing" most visual tokens. In contrast, our robust truncation (red histogram) yields a healthier, more spread-out distribution, ensuring that valid visual signals retain meaningful magnitudes.

As illustrated in Figure 7, informative regions (e.g., the contour of the motorcycle, the body of the giraffe, or the flock of birds) are suppressed and appear dark blue due to the dominance of a few bright spots in the “Before Truncation” maps. However, after applying our Robust Truncation, these structural details are vividly recovered. As shown in the “After Truncation” column, the heatmaps accurately highlight the semantic objects, confirming that our strategy is essential for preserving intrinsic visual information in the presence of attention artifacts.

D.2 Visualizations on Retained Tokens

To further demonstrate the precision of CrisPrune, we visualize the retained visual tokens on the COCO images using instructions from the POPE benchmark in Figure 9. The instructions follow the template “Is there a *{object}*?”, requiring the model to verify the existence of specific entities. CrisPrune accurately retains visual tokens corresponding to the queried objects (highlighted in red), even when they occupy a minimal spatial area or blend into complex backgrounds. Crucially, our method does not isolate the object in a void; it simultaneously preserves sparse yet informative background tokens. This confirms that CrisPrune effectively balances specific instruction alignment with intrinsic visual saliency.

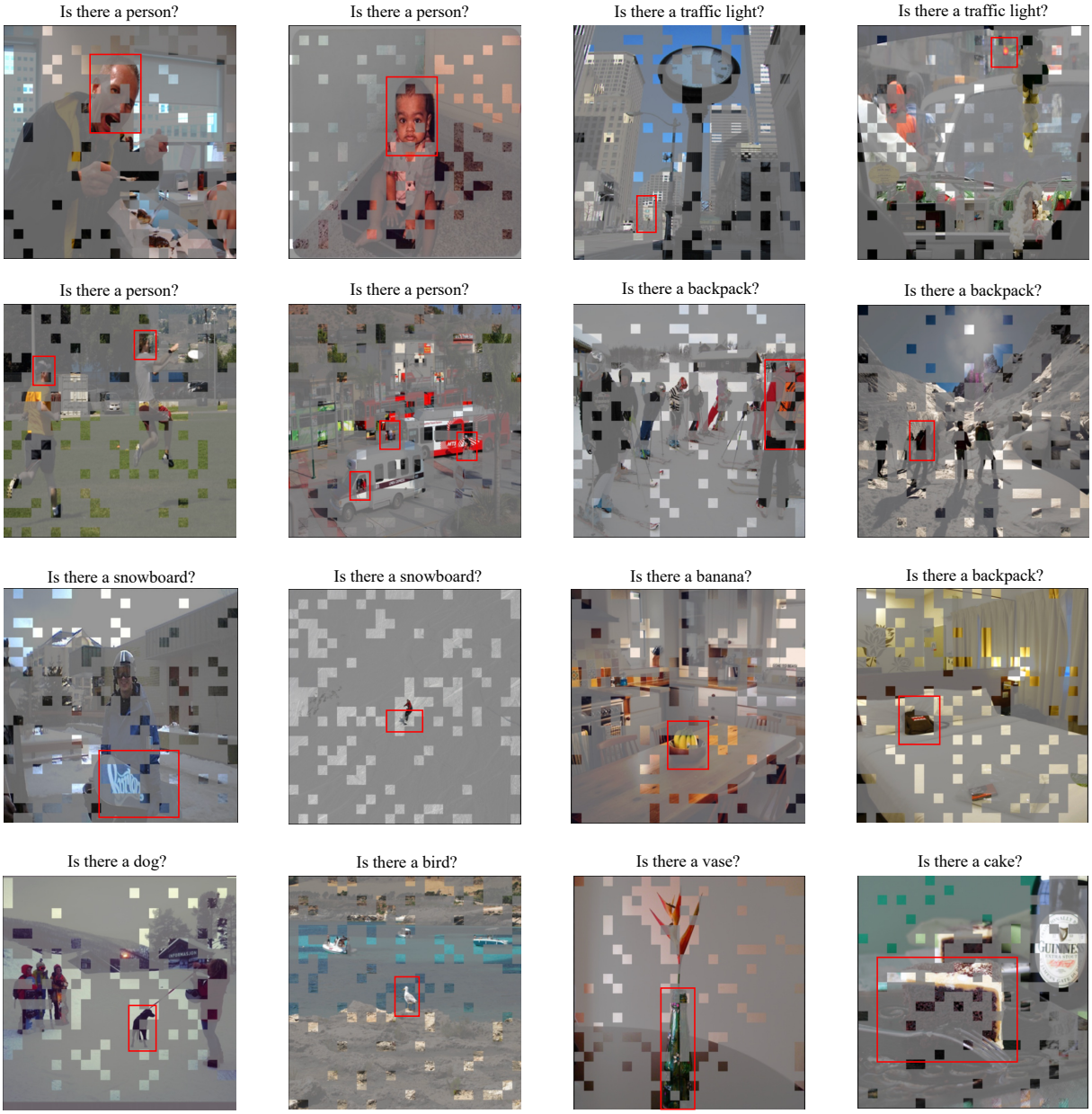

Figure 9: **Visualization of retained visual tokens by CrisPrune.** The images are chosen from the COCO dataset.