# Decoupling Spatial and Semantic Token Compression for Vision-Language Model Acceleration

**Anonymous Authors**[1]

## Abstract

Vision-Language Models (VLMs) face massive inference overhead from extensive visual tokens. Existing Top-$K$ pruning methods mitigate this but suffer from severe spatial bias, information redundancy, and crucial context loss. To address this, we propose TokenNMS, a training-free two-stage framework that reframes token reduction as deterministic feature-space Non-Maximum Suppression (NMS). TokenNMS seamlessly bridges query-agnostic spatial pruning with query-aware semantic filtering, enforcing similarity constraints to penalize semantic overlap. Extensive experiments demonstrate our approach effectively preserves spatially diverse representations while accelerating inference across diverse VLMs.

## 1. Introduction

Vision-Language Models (VLMs) (Liu et al., 2023; 2024a; Bai et al., 2025) encode high-resolution images into extensive visual token sequences (Dosovitskiy et al., 2020), causing excessive computational overhead due to the quadratic complexity of LLM attention mechanisms. Consequently, visual token pruning has emerged as a critical research area to mitigate inference costs.

Existing pruning methods (Chen et al., 2024; Zhang et al., 2024; 2025a; Xing et al., 2024; Shang et al., 2025) predominantly rely on Top-$K$ selection strategies based on predefined importance criteria. Despite introducing various compensatory mechanisms or dynamic retention ratios, these approaches inherently depend on strict, score-centric ranking. However, Top-$K$ selection suffers from critical limitations: (i) *spatial bias* towards salient objects (Endo et al., 2025) (Figure 1 (a)); (ii) *high redundancy* with overlapping features (Figure 2) ignoring decisive small objects

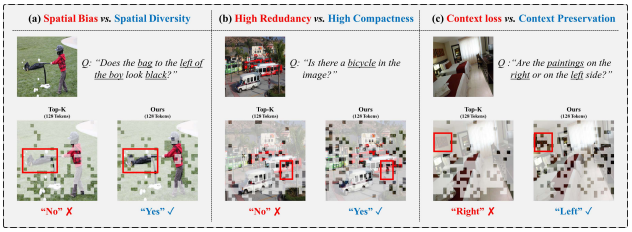


*Figure 1.* Illustration of comparison between conventional Top-$K$ approach and our TokenNMS on visual token retention on LLaVA-1.5-7B.

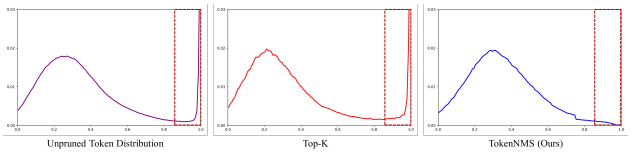


*Figure 2.* Cosine similarity distribution among tokens. Red dashed boxes indicate highly redundant tokens with high similarity scores.

(Alvar et al., 2025) (Figure 1 (b)); and (iii) *context loss* by eliminating necessary background cues (Huang et al., 2026) (Figure 1 (c)). This winner-takes-all distribution deprives the model of the global context required for complex reasoning.

To address these limitations, we propose TokenNMS, a training-free two-stage pruning framework. We reframe visual pruning as deterministic Non-Maximum Suppression (NMS) within the semantic feature space, enforcing distance constraints to prevent local clustering. Unlike Top-$K$ methods, TokenNMS penalizes semantic overlap to maintain a comprehensive, spatially diverse representation, as illustrated in Figure 2. It seamlessly bridges query-agnostic spatial pruning, ensuring visual diversity via NMS, with query-aware semantic filtering that selectively isolates task-relevant tokens as shown in Figure 1.

Extensive experiments across VLMs (e.g., LLaVA (Liu et al., 2023; 2024a), Qwen2.5-VL (Bai et al., 2025)) validate our approach. For instance, TokenNMS preserves 98.45% of unpruned LLaVA-NeXT-7B (Liu et al., 2024a) performance while reducing FLOPs by 90%. In summary, our contributions are three-fold: (1) we introduce a deterministic feature-space NMS strategy to suppress redundancy and enhance spatial diversity; (2) we propose TokenNMS,

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

decoupling query-agnostic spatial pruning and query-aware semantic filtering; and (3) we achieve an optimal efficiency-accuracy trade-off across diverse VQA benchmarks with minimal performance degradation.

## 2. Methodology

In this section, we first formulate the computational bottleneck of VLMs and expose the structural limitations of standard Top-$K$ selection approaches. Then we introduce a feature-space deterministic Non-Maximum Suppression (NMS) strategy, which aims to retain optimal spatial diversity and prevent information redundancy in high-dimensional embedding space. Based on this formulation, we propose TokenNMS, a hierarchical two-stage pruning framework. Lastly, we present the two primary components of our framework. The first stage performs query-agnostic pruning to preserve global context through visual saliency, which is then followed by a query-aware semantic filtering stage designed to extract task-relevant tokens via text-visual alignment.

### 2.1. Preliminaries

In Vision-Language Models (VLMs), encoding high-resolution images generates extensive visual token sequences, which introduces a severe computational bottleneck due to the quadratic complexity of LLM self-attention (Liu et al., 2023; Vaswani et al., 2017). To mitigate this, conventional token pruning employs a Top-$K$ selection strategy that retains a subset of tokens strictly based on individual importance scores (e.g., attention weights). However, this score-centric formulation intrinsically ignores pairwise semantic relationships between tokens (Wen et al., 2025; Alvar et al., 2025; Zhang et al., 2025b). Consequently, when visual saliency is highly concentrated, multiple tokens encoding nearly identical features dominate the top ranks, causing severe spatial collapse and information redundancy. This critical limitation necessitates a pruning metric that simultaneously evaluates individual saliency and mutual diversity, naturally leading to our proposed feature-space Non-Maximum Suppression (NMS) strategy.

### 2.2. Feature-Space Deterministic Non-Maximum Suppression

To mitigate the spatial bias and information redundancy of Top-$K$ pruning, we propose a feature-space NMS strategy. Inspired by the minimum distance constraint of Blue Noise sampling (Cook, 1986; Ulichney, 2002), we enforce a deterministic spatial repulsion principle within the high-dimensional semantic space of VLMs, avoiding the inference instability of stochastic processes (Bridson, 2007).

We replace physical distance with semantic correlation, quantified by the cosine similarity between token embeddings $x_i$ and $x_j$:

$$\text{sim}(x_i, x_j) = \frac{x_i^\top x_j}{\|x_i\|_2 \|x_j\|_2}. \tag{1}$$

To enforce a minimum semantic distance, we introduce a similarity threshold $\tau$. Tokens are considered to occupy highly overlapping semantic regions if their similarity exceeds $\tau$.

We iteratively construct a sparse, spatially diverse token set $\mathcal{V}$, initialized as $\mathcal{V} = \emptyset$. Traversing the candidate tokens based on their importance-based permutation $\pi$ (Section 2.1), each token $x_{\pi(i)}$ is evaluated against all currently selected tokens in $\mathcal{V}$. The candidate is appended to $\mathcal{V}$ if and only if it satisfies the strict distance constraint:

$$\max_{v \in \mathcal{V}} \text{sim}(x_{\pi(i)}, v) \leq \tau. \tag{2}$$

Conversely, if $\max_{v \in \mathcal{V}} \text{sim}(x_{\pi(i)}, v) > \tau$, the token $x_{\pi(i)}$ is suppressed as redundant. By greedily selecting salient tokens while rigidly penalizing semantic overlap, this deterministic NMS process yields a highly informative and structurally diverse visual token subset.

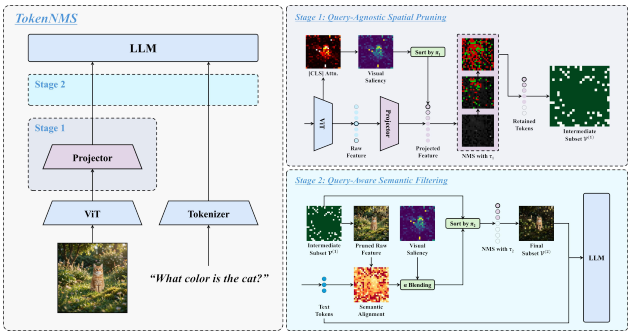


*Figure 3.* Illustration of TokenNMS with 2-stage visual token pruning framework.

### 2.3. TokenNMS

Our TokenNMS framework (Figure 3) seamlessly bridges query-agnostic spatial pruning and query-aware semantic filtering, effectively decoupling global context preservation from task-relevant feature isolation.

**Query-Agnostic Spatial Pruning.** To conservatively compress the massive visual token sequence into a structurally diverse subset of budget $K_1$, we extract visual saliency directly from the ViT `[CLS]` token self-attention prior to the LLM projection. Let $A \in \mathbb{R}^N$ denote the average attention map. To stabilize scales, we apply 99th-percentile clipping (Darcet et al., 2023) and min-max normalize it to obtain the saliency score $s_i$ for each projected token $x_i^{\text{proj}} \in \mathbb{R}^D$:

$$s_i = S_{\text{vis}}(x_i^{\text{proj}}) = \frac{A_i - \min(A)}{\max(A) - \min(A)}. \tag{3}$$

We sort the projected token set $X_{\text{proj}}$ by $s_i$ to obtain permutation $\pi_1$. Guided by $\pi_1$, we iteratively construct the intermediate subset $\mathcal{V}^{(1)}$, initialized empty. A candidate token $x^{\text{proj}}_{\pi_1(i)}$ is admitted to $\mathcal{V}^{(1)}$ only if its maximum cosine similarity to previously selected tokens satisfies the spatial threshold $\tau_1$:

$$\max_{v \in \mathcal{V}^{(1)}} \text{sim}(x^{\text{proj}}_{\pi_1(i)}, v) \leq \tau_1. \tag{4}$$

This NMS process in the projected LLM embedding space preserves highly attended regions while filtering out redundant background textures.

**Query-Aware Semantic Filtering.** The second stage refines $\mathcal{V}^{(1)}$ to a final budget $K_2 \ll K_1$ by isolating query-relevant features. To compute semantic alignment $r_j$, we utilize the raw visual feature $e_j \in \mathbb{R}^C$ prior to projection and the query embeddings $Q = \{q_1, \ldots, q_M\} \in \mathbb{R}^{M \times C}$ in the aligned contrastive space (Radford et al., 2021). We use average pooling to preserve holistic context:

$$r_j = S_{\text{sem}}(e_j, Q) = \frac{1}{M} \sum_{m=1}^{M} \frac{e_j^\top q_m}{\|e_j\|_2 \|q_m\|_2}. \tag{5}$$

To balance semantic alignment with spatial perception, we employ a joint importance score:

$$S_{\text{joint}}(v_j) = \alpha \cdot s_j + (1 - \alpha) \cdot \tilde{r}_j, \tag{6}$$

where $\tilde{r}_j$ is the min-max normalized semantic alignment and $\alpha$ controls the balance. We sort $\mathcal{V}^{(1)}$ by $S_{\text{joint}}$ to obtain permutation $\pi_2$.

Crucially, to prevent semantic collapse—where spatially distinct objects sharing a high-level concept mutually suppress each other—we decouple the suppression criterion from the selection priority. We impose the NMS constraint in the pre-projection visual space ($\mathbb{R}^C$). Evaluating candidate $v_{\pi_2(j)}$ against the selected set $\mathcal{V}^{(2)}$, we apply the decoupled criterion using raw features $e_{\pi_2(j)}$ and $e_u$:

$$\max_{u \in \mathcal{V}^{(2)}} \text{sim}_{\text{vis}}(e_{\pi_2(j)}, e_u) \leq \tau_2, \tag{7}$$

where $\text{sim}_{\text{vis}}$ is the cosine similarity in dimension $C$ and $\tau_2$ is the decoupled threshold. By ranking semantically but suppressing visually, TokenNMS extracts a compact, query-grounded representation while maintaining spatial multiplicity.

# 3. Experiments

## 3.1. Experimental Setup

We evaluate TokenNMS on LLaVA-1.5-7B (Liu et al., 2024a), LLaVA-NeXT-7B (Liu et al., 2024b), and Qwen2.5-VL-7B (Bai et al., 2025) across diverse multimodal benchmarks, including VQAv2 (Goyal et al., 2017), GQA (Hudson & Manning, 2019), TextVQA (Singh et al., 2019), and POPE (Li et al., 2023). Token deficits from NMS suppression are deterministically padded with top unselected candidates to support batched inference. For hyperparameters, Stage 1 conservatively retains ∼88% of tokens ($\tau_1 = 0.6$) to eliminate only highly redundant backgrounds. Stage 2 employs $\alpha = 0.5$ to balance visual structure with query relevance, and $\tau_2 = 0.8$ to prevent over-suppressing query-critical features. Additional results and details can be found in the Appendix.

*Table 1.* Performance comparison of different visual token reduction methods on LLaVA-1.5-7B across multiple VQA benchmarks. **Rel.** indicates the relative performance preserved compared to the upper bound.

| Method | VQA$^{V2}$ | GQA | SQA$^{IMG}$ | VQA$^{Text}$ | POPE | MME | MMB | MMVet | Rel. |
|---|---|---|---|---|---|---|---|---|---|
| *Upper Bound, All 576 Tokens* (100%) | | | | | | | | | |
| LLaVA-1.5-7B | 78.5 | 61.9 | 69.5 | 58.2 | 85.9 | 1506.5 | 64.7 | 31.3 | 100.00% |
| *Retain 192 Tokens* (↓ 66.7%) | | | | | | | | | |
| FastV (Chen et al., 2024) | 67.1 | 52.7 | 67.9 | 52.5 | 64.8 | 1398.4 | 61.2 | 27.7 | 88.73% |
| PyramidDrop (Xing et al., 2024) | 74.9 | 57.1 | 70.2 | 56.1 | 82.3 | 1465.9 | 63.2 | 30.5 | 96.66% |
| SparseVLM (Zhang et al., 2024) | 75.6 | 57.6 | 69.1 | 56.1 | 83.6 | 1452.9 | 62.5 | 31.5 | 97.02% |
| VisionZip (Yang et al., 2025) | 76.8 | 59.3 | 68.9 | 57.3 | 85.3 | 1438.6 | 63.0 | 31.7 | 98.08% |
| CDPruner (Zhang et al., 2025b) | 77.2 | 60.3 | 68.8 | 57.3 | 87.3 | 1446.2 | 63.2 | 30.5 | 98.24% |
| **TokenNMS (Ours)** | 77.3 | 60.5 | 69.1 | 57.3 | 87.5 | 1462.1 | 63.6 | 32.1 | **99.23%** |
| *Retain 128 Tokens* (↓ 77.8%) | | | | | | | | | |
| FastV (Chen et al., 2024) | 71.0 | 54.0 | 69.2 | 56.4 | 68.2 | 1368.9 | 63.0 | 27.0 | 91.01% |
| PyramidDrop (Xing et al., 2024) | 74.3 | 57.1 | 70.1 | 56.7 | 77.5 | 1444.1 | 62.3 | 27.6 | 94.47% |
| SparseVLM (Zhang et al., 2024) | 75.1 | 57.3 | 69.0 | 56.3 | 83.1 | 1399.3 | 62.6 | 29.7 | 95.69% |
| PruMerge+ (Shang et al., 2025) | 75.0 | 58.2 | 69.1 | 54.0 | 83.1 | 1408.1 | 61.8 | 30.4 | 95.58% |
| VisionZip (Yang et al., 2025) | 75.6 | 57.6 | 68.7 | 56.9 | 83.3 | 1436.9 | 62.1 | 31.6 | 96.91% |
| DivPrune (Alvar et al., 2025) | 76.0 | 59.4 | 68.6 | 55.9 | 87.0 | 1405.1 | 61.5 | 30.6 | 96.86% |
| CDPruner (Zhang et al., 2025b) | 76.6 | 59.9 | 69.0 | 56.2 | 87.7 | 1431.4 | 63.1 | 32.8 | 98.70% |
| **TokenNMS (Ours)** | 76.7 | 60.0 | 68.5 | 56.6 | 87.3 | 1461.3 | 63.1 | 32.4 | **98.77%** |
| *Retain 64 Tokens* (↓ 88.9%) | | | | | | | | | |
| FastV (Chen et al., 2024) | 55.9 | 46.0 | 70.1 | 51.6 | 35.5 | 973.5 | 50.1 | 18.9 | 72.35% |
| PyramidDrop (Xing et al., 2024) | 56.3 | 46.1 | 68.8 | 49.2 | 40.8 | 982.2 | 48.0 | 17.7 | 71.64% |
| SparseVLM (Zhang et al., 2024) | 66.9 | 52.0 | 59.2 | 52.1 | 69.7 | 1190.4 | 58.3 | 24.4 | 84.02% |
| PruMerge+ (Shang et al., 2025) | 71.3 | 55.4 | 69.5 | 52.0 | 75.7 | 1316.8 | 59.6 | 28.0 | 90.85% |
| VisionZip (Yang et al., 2025) | 72.4 | 55.1 | 69.0 | 55.5 | 77.0 | 1365.2 | 60.1 | 29.4 | 92.87% |
| DivPrune (Alvar et al., 2025) | 74.1 | 57.5 | 68.0 | 54.5 | 85.5 | 1334.7 | 60.1 | 28.1 | 93.70% |
| CDPruner (Zhang et al., 2025b) | 75.4 | 58.6 | 68.1 | 55.3 | 87.5 | 1415.1 | 61.1 | 30.5 | 96.42% |
| **TokenNMS (Ours)** | 75.6 | 58.9 | 69.3 | 55.9 | 87.4 | 1423.7 | 61.4 | 30.9 | **97.14%** |

## 3.2. Main Results

**Performance on LLaVA-1.5-7B.** Table 1 demonstrates TokenNMS's superiority across varying reduction rates. When retaining 192 tokens, it preserves 99.23% of vanilla performance. Under extreme compression (64 tokens), TokenNMS robustly maintains 97.14%, effectively preserving crucial spatial context compared to existing Top-$K$ baselines.

**Scalability to High-Resolution VLMs.** Table 2 highlights our scalability to massive token sequences via LLaVA-NeXT-7B (Liu et al., 2024b). At a 640-token budget, TokenNMS achieves 100.78% relative performance, surpassing the unpruned baseline by successfully filtering redundant visual noise. Remarkably, even at an extreme 94.4% compression rate, it preserves 96.71% of the original performance. Evaluations on Qwen2.5-VL (Bai et al., 2025) further confirm our framework's strong generalizability across diverse architectures and text-dense benchmarks.

Table 3 details results on Qwen2.5-VL-7B. TokenNMS achieves the highest relative performance at 512 (98.98%) and 256 (94.02%) token budgets, outperforming all baselines. Even under extreme compression (128 tokens), it

*Table 2.* Performance comparison of different visual token reduction methods on LLaVA-NeXT-7B across multiple VQA benchmarks. **Rel.** indicates the relative performance preserved compared to the upper bound.

| Method | $VQA^{V2}$ | GQA | $SQA^{IMG}$ | $VQA^{Text}$ | POPE | MME | MMB | MMVet | Rel. |
|---|---|---|---|---|---|---|---|---|---|
| *Upper Bound, All 2880 Tokens* (100%) | | | | | | | | | |
| LLaVA-NeXT-7B | 81.3 | 62.5 | 67.5 | 60.3 | 86.8 | 1511.8 | 65.8 | 40.0 | 100.00% |
| *Retain 640 Tokens* (↓ **77.8**%) | | | | | | | | | |
| FastV (Chen et al., 2024) | 77.0 | 58.9 | 67.4 | 58.1 | 79.5 | 1412.6 | 63.1 | 39.5 | 95.60% |
| PyramidDrop (Xing et al., 2024) | 79.1 | 60.0 | 66.7 | 57.8 | 83.8 | 1475.9 | 64.1 | 36.7 | 96.41% |
| SparseVLM (Zhang et al., 2024) | 79.2 | 61.2 | 67.6 | 59.7 | 85.3 | 1456.8 | 65.9 | 36.1 | 97.44% |
| PruMerge+ (Shang et al., 2025) | 78.2 | 60.8 | 67.8 | 54.9 | 85.3 | 1480.2 | 64.6 | 32.7 | 95.13% |
| VisionZip (Yang et al., 2025) | 79.1 | 61.2 | 68.1 | 59.9 | 86.0 | 1493.4 | 65.8 | 38.9 | 98.82% |
| DivPrune (Alvar et al., 2025) | 79.3 | 61.9 | 67.8 | 57.0 | 86.9 | 1469.7 | 65.8 | 38.0 | 97.99% |
| CDPruner (Zhang et al., 2025b) | 79.9 | 62.6 | 67.9 | 58.4 | 87.3 | 1474.5 | 66.3 | 41.9 | 99.94% |
| **TokenNMS (Ours)** | 80.1 | 65.5 | 68.1 | 59.8 | 87.5 | 1489.5 | 66.3 | 41.1 | **100.78%** |
| *Retain 320 Tokens* (↓ **88.9**%) | | | | | | | | | |
| FastV (Chen et al., 2024) | 61.5 | 49.8 | 66.6 | 52.2 | 49.5 | 1099.0 | 53.4 | 20.0 | 75.18% |
| PyramidDrop (Xing et al., 2024) | 66.8 | 50.4 | 66.7 | 49.0 | 60.8 | 1171.5 | 55.5 | 24.0 | 79.35% |
| SparseVLM (Zhang et al., 2024) | 74.6 | 57.9 | 67.2 | 56.5 | 76.9 | 1386.1 | 63.1 | 32.8 | 91.98% |
| PruMerge+ (Shang et al., 2025) | 75.3 | 58.8 | 68.1 | 54.0 | 79.5 | 1444.3 | 63.0 | 31.4 | 92.31% |
| VisionZip (Yang et al., 2025) | 76.2 | 58.9 | 67.5 | 58.8 | 82.3 | 1397.1 | 63.3 | 35.8 | 94.80% |
| DivPrune (Alvar et al., 2025) | 77.2 | 61.1 | 67.7 | 56.2 | 84.7 | 1423.3 | 63.9 | 34.8 | 95.26% |
| CDPruner (Zhang et al., 2025b) | 78.4 | 61.6 | 67.8 | 57.4 | 87.2 | 1453.0 | 65.5 | 37.9 | 97.69% |
| **TokenNMS (Ours)** | 78.8 | 61.3 | 67.9 | 57.8 | 87.3 | 1500.1 | 65.7 | 38.6 | **98.45%** |
| *Retain 160 Tokens* (↓ **94.4**%) | | | | | | | | | |
| PruMerge+ (Shang et al., 2025) | 70.5 | 56.2 | 66.9 | 50.3 | 71.1 | 1289.6 | 58.0 | 29.3 | 85.97% |
| VisionZip (Yang et al., 2025) | 71.4 | 55.2 | 67.9 | 55.0 | 74.9 | 1327.8 | 58.6 | 32.3 | 88.98% |
| DivPrune (Alvar et al., 2025) | 75.0 | 59.3 | 67.1 | 54.1 | 80.0 | 1356.6 | 62.9 | 32.0 | 91.72% |
| CDPruner (Zhang et al., 2025b) | 76.7 | 60.8 | 67.5 | 55.4 | 86.8 | 1425.3 | 64.2 | 36.2 | 95.73% |
| **TokenNMS (Ours)** | 77.0 | 60.6 | 67.6 | 57.2 | 87.1 | 1458.2 | 65.1 | 36.5 | **96.71%** |

*Table 3.* Performance comparison of different visual token reduction methods on Qwen2.5-VL-7B across multiple VQA benchmarks. **Rel.** indicates the relative performance preserved compared to the upper bound. The best and second-best results are highlighted in **bold** and underlined, respectively.

| Method | ChartQA | AI2D | OCRBench | HallBench | MME | MMB | Rel. |
|---|---|---|---|---|---|---|---|
| *Upper Bound, All 1296 Tokens* (100%) | | | | | | | |
| Qwen2.5-VL-7B | 86.1 | 80.4 | 863 | 64.8 | 2304 | 82.8 | 100.00% |
| *Retain 512 Tokens* (↓ **60.5**%) | | | | | | | |
| FastV (Chen et al., 2024) | 82.2 | 78.8 | 815 | 62.1 | 2317 | 82.0 | 97.62% |
| DivPrune (Alvar et al., 2025) | 79.6 | 78.6 | 800 | 60.9 | 2279 | 81.6 | 96.34% |
| CDPruner (Zhang et al., 2025b) | 82.8 | 78.9 | 827 | 62.3 | 2327 | 82.2 | 98.08% |
| **TokenNMS (Ours)** | 83.6 | 79.9 | 827 | 64.0 | 2313 | 83.9 | **98.97%** |
| *Retain 256 Tokens* (↓ **80.2**%) | | | | | | | |
| FastV (Chen et al., 2024) | 70.9 | 76.2 | 703 | 59.6 | 2238 | 79.6 | 91.98% |
| DivPrune (Alvar et al., 2025) | 65.1 | 76.5 | 692 | 57.4 | 2184 | 80.0 | 90.13% |
| CDPruner (Zhang et al., 2025b) | 73.0 | 77.5 | 749 | 59.7 | 2245 | 80.9 | 93.61% |
| **TokenNMS (Ours)** | 73.2 | 78.1 | 727 | 61.3 | 2252 | 82.3 | **94.02%** |
| *Retain 128 Tokens* (↓ **90.1**%) | | | | | | | |
| FastV (Chen et al., 2024) | 52.2 | 71.4 | 531 | 49.2 | 2008 | 72.9 | 80.30% |
| DivPrune (Alvar et al., 2025) | 50.4 | 72.1 | 549 | 53.6 | 2108 | 77.8 | 82.86% |
| CDPruner (Zhang et al., 2025b) | 59.2 | 74.0 | 632 | 55.4 | 2127 | 76.2 | **86.26%** |
| **TokenNMS (Ours)** | 56.4 | 75.0 | 554 | 55.7 | 2087 | 80.4 | 85.23% |

maintains a highly competitive 85.23%, securing top scores on tasks like AI2D, HallusionBench, and MMBench. This confirms TokenNMS's robust generalizability across diverse architectures.

### 3.3. Efficiency Analysis

We evaluate computational efficiency on LLaVA-NeXT-7B using a single RTX 4090 GPU and the POPE benchmark. We adopted POPE for its uniform prompt lengths and straightforward single-prefill, single-decode inference structure. The TFLOPs were computed by summing the floating-point operations for each transformer computation. The prefill time (*i.e.*, Time-To-First-Token) was measured by generating a single token and counting the elapsed time from input processing to the first output token. The decode time was computed by subtracting the prefill time from the total generation time and dividing by the number of subsequently generated tokens, yielding the average latency per token during autoregressive decoding. The KV cache size was calculated based on the number of layers, key-value heads, sequence length, head dimension, and data precision. As detailed in Table 4, retaining 320 tokens with TokenNMS reduces computational cost (TFLOPs) by roughly 10x and KV cache size by nearly 90% compared to the vanilla baseline. Furthermore, TokenNMS achieves the lowest TFLOPs and fastest decode latency among existing methods. While SparseVLM yields a slightly faster prefill time, it suffers from severe performance degradation and higher memory usage. Ultimately, these results demonstrate that TokenNMS maximizes inference acceleration and token informativeness with minimal overhead.

*Table 4.* Efficiency and performance comparison across various pruning methods on POPE. We report FLOPs, prefill/decode time, KV cache size, GPU memory footprint, and the resulting F1 score on LLaVA-NeXT-7B.

| Method | # Token | TFLOPs | Prefill Time (ms) | Decode Time (ms/token) | KV Cache (MB) | GPU Memory (GB) | Score (F1) |
|---|---|---|---|---|---|---|---|
| LLaVA-NeXT-7B | 2880 | 41.8 | 393 | 13 | 1543 | 18.5 | 86.8 |
| PyramidDrop (Xing et al., 2024) | 320 | 4.7 | 168 | 10 | 190 | **15.8** | 60.8 |
| SparseVLM (Zhang et al., 2024) | 320 | 4.8 | 163 | 10 | 193 | 20.6 | 76.9 |
| VisionZip (Yang et al., 2025) | 320 | 5 | **105** | 11 | **167** | 16.2 | 82.3 |
| CDPruner (Zhang et al., 2025b) | 320 | **4.2** | 162 | 10 | 168 | 16.4 | 87.2 |
| **TokenNMS (Ours)** | 320 | **4.2** | 165 | **9** | **167** | 16.4 | **87.3** |

## 4. Conclusion

In this paper, we introduce TokenNMS, a training-free two-stage pruning framework for Vision-Language Model (VLM) inference acceleration. By reframing token selection through a feature-space Non-Maximum Suppression (NMS) strategy, we effectively mitigated the spatial bias and redundancy inherent in conventional Top-$K$ methods. Integrating query-agnostic spatial pruning with query-aware semantic filtering enables massive token reduction without sacrificing crucial visual context. In addition, decoupling selection priority from the suppression criterion allows the model to rank tokens by task relevance while ensuring spatial diversity through raw visual feature constraints. Extensive experiments on various VLMs and benchmarks confirm that TokenNMS achieves highly competitive hardware efficiency and maintains exceptional reasoning performance under extreme compression, providing a robust solution for efficient VLM deployment.

## Impact Statement

This paper presents work aimed at advancing the efficiency of Vision-Language Models. By reducing the computational overhead of multimodal inference, our approach has the potential to lower the energy consumption and carbon footprint of deploying large models, while democratizing access to advanced AI on resource-constrained devices.

## References


Agrawal, H., Desai, K., Wang, Y., Chen, X., Jain, R., Johnson, M., Batra, D., Parikh, D., Lee, S., and Anderson, P. Nocaps: Novel object captioning at scale. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 8948–8957, 2019.

Alvar, S. R., Singh, G., Akbari, M., and Zhang, Y. Divprune: Diversity-based visual token pruning for large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 9392–9401, 2025.

Arif, K. H. I., Yoon, J., Nikolopoulos, D. S., Vandierendonck, H., John, D., and Ji, B. Hired: Attention-guided token dropping for efficient inference of high-resolution vision-language models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 1773–1781, 2025.

Bai, J., Bai, S., Chu, Y., Cui, Z., Dang, K., Deng, X., Fan, Y., Ge, W., Han, Y., Huang, F., et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., and Lin, J. Qwen2.5-vl technical report. *ArXiv*, abs/2502.13923, 2025.

Bolya, D., Fu, C.-Y., Dai, X., Zhang, P., Feichtenhofer, C., and Hoffman, J. Token merging: Your vit but faster. *arXiv preprint arXiv:2210.09461*, 2022.

Bridson, R. Fast poisson disk sampling in arbitrary dimensions. *SIGGRAPH sketches*, 10(1):1, 2007.

Chen, L., Zhao, H., Liu, T., Bai, S., Lin, J., Zhou, C., and Chang, B. An image is worth 1/2 tokens after layer 2: Plug-and-play inference acceleration for large vision-language models. In *European Conference on Computer Vision*, pp. 19–35. Springer, 2024.

Chen, X., Fang, H., Lin, T.-Y., Vedantam, R., Gupta, S., Dollár, P., and Zitnick, C. L. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

Cook, R. L. Stochastic sampling in computer graphics. *ACM Transactions on Graphics (TOG)*, 5(1):51–72, 1986.

Darcet, T., Oquab, M., Mairal, J., and Bojanowski, P. Vision transformers need registers. *arXiv preprint arXiv:2309.16588*, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Endo, M., Wang, X., and Yeung-Levy, S. Feather the throttle: Revisiting visual token pruning for vision-language model acceleration. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22826–22835, 2025.

Fu, C., Chen, P., Shen, Y., Qin, Y., Zhang, M., Lin, X., Yang, J., Zheng, X., Li, K., Sun, X., et al. Mme: A comprehensive evaluation benchmark for multimodal large language models. *arXiv preprint arXiv:2306.13394*, 2023.

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

Guan, T., Liu, F., Wu, X., Xian, R., Li, Z., Liu, X., Wang, X., Chen, L., Huang, F., Yacoob, Y., et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14375–14385, 2024.

Gurari, D., Li, Q., Stangl, A. J., Guo, A., Lin, C., Grauman, K., Luo, J., and Bigham, J. P. Vizwiz grand challenge: Answering visual questions from blind people. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 3608–3617, 2018.

Huang, Y., Ma, F., Shao, Y., Guo, J., Yu, Z., Cui, L., and Tian, Q. N\" uwa: Mending the spatial integrity torn by vlm token pruning. *arXiv preprint arXiv:2602.02951*, 2026.

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

Jin, P., Takanobu, R., Zhang, W., Cao, X., and Yuan, L. Chat-univi: Unified visual representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 13700–13710, 2024.

Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., and Farhadi, A. A diagram is worth a dozen images. In *European conference on computer vision*, pp. 235–251. Springer, 2016.

Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., and Qiao, Y. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68 (10):200102, 2025.

Li, Y., Du, Y., Zhou, K., Wang, J., Zhao, W. X., and Wen, J.-R. Evaluating object hallucination in large vision-language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pp. 292–305, 2023.

Li, Y., Huang, Y., Yang, B., Venkitesh, B., Locatelli, A., Ye, H., Cai, T., Lewis, P., and Chen, D. Snapkv: Llm knows what you are looking for before generation. *Advances in Neural Information Processing Systems*, 37:22947–22970, 2024.

Liang, Y., Ge, C., Tong, Z., Song, Y., Wang, J., and Xie, P. Not all patches are what you need: Expediting vision transformers via token reorganizations. *arXiv preprint arXiv:2202.07800*, 2022.

Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., and Yuan, L. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., and Zitnick, C. L. Microsoft coco: Common objects in context. In *European conference on computer vision*, pp. 740–755. Springer, 2014.

Liu, H., Li, C., Wu, Q., and Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Liu, H., Li, C., Li, Y., and Lee, Y. J. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 26296–26306, 2024a.

Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., and Lee, Y. J. Llavanext: Improved reasoning, ocr, and world knowledge, 2024b.

Liu, Y., Duan, H., Zhang, Y., Li, B., Zhang, S., Zhao, W., Yuan, Y., Wang, J., He, C., Liu, Z., et al. Mmbench: Is your multi-modal model an all-around player? In *European conference on computer vision*, pp. 216–233. Springer, 2024c.

Liu, Y., Li, Z., Huang, M., Yang, B., Yu, W., Li, C., Yin, X.-C., Liu, C.-L., Jin, L., and Bai, X. Ocrbench: on the hidden mystery of ocr in large multimodal models. *Science China Information Sciences*, 67(12):220102, 2024d.

Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.-W., Zhu, S.-C., Tafjord, O., Clark, P., and Kalyan, A. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in neural information processing systems*, 35:2507–2521, 2022.

Masry, A., Do, X. L., Tan, J. Q., Joty, S., and Hoque, E. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the association for computational linguistics: ACL 2022*, pp. 2263–2279, 2022.

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Rao, Y., Zhao, W., Liu, B., Lu, J., Zhou, J., and Hsieh, C.-J. Dynamicvit: Efficient vision transformers with dynamic token sparsification. *Advances in neural information processing systems*, 34:13937–13949, 2021.

Shang, Y., Cai, M., Xu, B., Lee, Y. J., and Yan, Y. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 22857–22867, 2025.

Singh, A., Natarajan, V., Shah, M., Jiang, Y., Chen, X., Batra, D., Parikh, D., and Rohrbach, M. Towards vqa models that can read. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 8317–8326, 2019.

Tay, Y., Dehghani, M., Bahri, D., and Metzler, D. Efficient transformers: A survey. *ACM Computing Surveys*, 55(6): 1–28, 2022.

Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.-A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al. Llama: Open and efficient foundation language models. arxiv 2023. *arXiv preprint arXiv:2302.13971*, 10, 2023a.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Ulichney, R. A. Dithering with blue noise. *Proceedings of the IEEE*, 76(1):56–79, 2002.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Wen, Z., Gao, Y., Wang, S., Zhang, J., Zhang, Q., Li, W., He, C., and Zhang, L. Stop looking for "important tokens" in multimodal language models: Duplication matters more. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 9972–9991, 2025.

Xiao, G., Tian, Y., Chen, B., Han, S., and Lewis, M. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*, 2023.

Xing, L., Huang, Q., Dong, X., Lu, J., Zhang, P., Zang, Y., Cao, Y., He, C., Wang, J., Wu, F., et al. Pyramiddrop: Accelerating your large vision-language models via pyramid visual redundancy reduction. *arXiv preprint arXiv:2410.17247*, 2024.

Yang, S., Chen, Y., Tian, Z., Wang, C., Li, J., Yu, B., and Jia, J. Visionzip: Longer is better but not necessary in vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 19792–19802, 2025.

Yu, W., Yang, Z., Li, L., Wang, J., Lin, K., Liu, Z., Wang, X., and Wang, L. Mm-vet: Evaluating large multimodal models for integrated capabilities. *arXiv preprint arXiv:2308.02490*, 2023.

Zhang, Q., Cheng, A., Lu, M., Zhang, R., Zhuo, Z., Cao, J., Guo, S., She, Q., and Zhang, S. Beyond text-visual attention: Exploiting visual cues for effective token pruning in vlms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 20857–20867, 2025a.

Zhang, Q., Liu, M., Li, L., Lu, M., Zhang, Y., Pan, J., She, Q., and Zhang, S. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *arXiv preprint arXiv:2506.10967*, 2025b.

Zhang, Y., Fan, C.-K., Ma, J., Zheng, W., Huang, T., Cheng, K., Gudovskiy, D., Okuno, T., Nakata, Y., Keutzer, K., et al. Sparsevlm: Visual token sparsification for efficient vision-language model inference. *arXiv preprint arXiv:2410.04417*, 2024.

Zhang, Z., Sheng, Y., Zhou, T., Chen, T., Zheng, L., Cai, R., Song, Z., Tian, Y., Ré, C., Barrett, C., et al. H2o: Heavy-hitter oracle for efficient generative inference of large language models. *Advances in Neural Information Processing Systems*, 36:34661–34710, 2023.

# A. Related Work

## A.1. Vision-Language Models

Vision-Language Models (VLMs) (Liu et al., 2023; 2024a; Wang et al., 2024) integrate a vision encoder with a large language model (LLM) (Touvron et al., 2023a;b; Bai et al., 2023) by aligning visual tokens to the LLM feature space. Through instruction tuning, VLMs serve as general-purpose reasoners for tasks ranging from descriptive generation (Chen et al., 2015; Agrawal et al., 2019) to fine-grained question answering (Goyal et al., 2017; Hudson & Manning, 2019; Gurari et al., 2018). However, handling high-resolution images or videos generates massive visual token sequences, causing a severe inference bottleneck. This token overload leads to several challenges, including the quadratic computational scaling of self-attention (Vaswani et al., 2017; Tay et al., 2022) and excessive memory pressure from KV caches (Xiao et al., 2023; Zhang et al., 2023; Li et al., 2024). Furthermore, the resulting increase in latency hinders the deployment of VLMs for long-context inputs (Lin et al., 2024; Li et al., 2025; Jin et al., 2024). Consequently, achieving efficient multimodal inference requires balancing the preservation of vital visual semantics with the reduction of token-dependent overhead, strongly motivating methods that prune or restructure visual tokens.

## A.2. Visual Token Compression

Visual token compression aims to shorten visual sequences during VLM inference to decrease computation and memory usage. Prior works (Rao et al., 2021; Liang et al., 2022) range from training-based compression, which requires architectural changes or fine-tuning, to training-free approaches. The latter are widely adopted for their seamless deployment. Token pruning seeks to retain only a highly informative subset of tokens, categorized mainly into attention-based and diversity-based methods. Attention-based pruning (Chen et al., 2024; Xing et al., 2024; Zhang et al., 2024; Arif et al., 2025) selects top-ranked tokens using vision or cross-attention scores. While straightforward, these scores can be unstable, and aggressive pruning often collapses retained tokens into limited spatial regions. Conversely, diversity-based pruning (Wen et al., 2025; Alvar et al., 2025) leverages feature similarity to preserve spatial and semantic diversity. However, depending exclusively on similarity often fails to capture instruction-specific relevance and incurs computational overhead. Hybrid methods (Yang et al., 2025; Shang et al., 2025; Zhang et al., 2025a;b) combine attention relevance with diversity constraints to balance query-awareness and redundancy, though this often necessitates complex selection mechanisms. Alternatively, token merging (Bolya et al., 2022; Huang et al., 2026) compresses the sequence by aggregating similar tokens. While it preserves global context without explicit token discarding, aggressive merging dilutes the fine-grained visual cues essential for text-dense or localization-centric tasks.

# B. Algorithmic Details of TokenNMS

To facilitate the reproducibility of our framework, we provide the detailed the algorithmic implementations of the proposed TokenNMS framework. Algorithm 1 formalizes the query-agnostic spatial pruning process. Given the projected visual tokens $X_{\text{proj}}$, we first compute the inherent visual saliency score $s_i$ for each token and sort the sequence in descending order. We then iteratively construct the intermediate subset $\mathcal{V}^{(1)}$ through a greedy selection process. A candidate token is appended to the subset only if its maximum cosine similarity to all previously selected tokens remains below the spatial threshold $\tau_1$. This Non-Maximum Suppression (NMS) effectively filters out highly redundant regions. The process terminates once the cardinality of the subset reaches the predefined budget $K_1$, yielding a structurally diverse visual representation.

Algorithm 2 details the query-aware semantic filtering stage. Taking the intermediate subset $\mathcal{V}^{(1)}$ as input, we first compute the semantic alignment score $r_j$ between each visual token and the text query $Q$. After applying min-max normalization to these scores, we calculate a joint importance metric $S_{\text{joint}}$ by blending the preserved visual saliency and the normalized semantic alignment using a weighting parameter $\alpha$. The tokens are subsequently re-sorted based on this joint score. To prevent semantic collapse, we apply a decoupled NMS strategy: a candidate token is admitted to the final set $\mathcal{V}^{(2)}$ only if its pre-projection visual feature $e_{\text{cand}}$ exhibits a similarity below the threshold $\tau_2$ when compared against the pre-projection features of the currently selected tokens. This loop concludes when the final target budget $K_2$ is met.

**Algorithm 1** TokenNMS: Query-Agnostic Spatial Pruning

**Input:** Projected visual tokens $X_{\text{proj}} = \{x_1^{\text{proj}}, \ldots, x_N^{\text{proj}}\}$, threshold $\tau_1$, token budget $K_1$
**Output:** Query-agnostic spatial pruned visual token set $\mathcal{V}^{(1)}$
Initialize query-agnostic token set $\mathcal{V}^{(1)} \leftarrow \emptyset$
For each token $x_i^{\text{proj}} \in X_{\text{proj}}$ do
    Compute visual saliency score $s_i = S_{\text{vis}}(x_i^{\text{proj}})$
End For
Sort $X_{\text{proj}}$ in descending order of $s_i$ to obtain permutation $\pi_1$
For $i = 1$ to $N$ do
    $x_{\text{cand}} \leftarrow x_{\pi_1(i)}^{\text{proj}}$
    If $\max_{v \in \mathcal{V}^{(1)}} \text{sim}_{\text{vis}}(x_{\text{cand}}, v) \leq \tau_1$ then
        $\mathcal{V}^{(1)} \leftarrow \mathcal{V}^{(1)} \cup \{x_{\text{cand}}\}$
    End If
    If $|\mathcal{V}^{(1)}| == K_1$ then
        break // Reach Stage 1 budget
    End If
End For
Return $\mathcal{V}^{(1)}$

# C. Experimental Setup Details

## C.1. Evaluated VLM Backbones

**LLaVA-1.5 (Liu et al., 2024a)** is a widely adopted Vision-Language Model (VLM) that integrates a vision encoder with a large language model through a simple projection layer. It operates on a fixed-resolution visual processing paradigm, uniformly generating a consistent number of visual tokens regardless of the original image dimensions. Due to its fixed token generation and standard architecture built upon a fine-tuned variant of the LLaMA family, it serves as a foundational baseline for evaluating the efficacy and computational trade-offs of various visual token pruning algorithms.

**LLaVA-NeXT (Liu et al., 2024b)** advances the baseline architecture by introducing a dynamic, high-resolution processing mechanism that significantly improves fine-grained visual comprehension. Instead of relying on a fixed-size input, it adopts a multi-patch strategy that yields a substantially larger and variable number of visual tokens, generating up to 2,880 tokens for a single input. This extended sequence length preserves detailed spatial information but inherently increases computational overhead. Consequently, LLaVA-NeXT provides a rigorous testbed for assessing the scalability and robustness of token reduction methods under heavy computational constraints.

**Qwen2.5-VL (Bai et al., 2025)** is a multimodal model designed with robust native support for high-resolution visual inputs and arbitrary aspect ratios. The architecture dynamically scales the number of generated visual tokens according to the intrinsic resolution and complexity of the input image. Furthermore, Qwen2.5-VL explicitly incorporates structured positional encodings, such as Rotary Position Embeddings (RoPE), to establish implicit visual coordinates essential for spatial reasoning. This capacity for native high-resolution processing and robust spatial representation makes it a critical backbone for evaluating spatially-aware and adaptive token pruning frameworks.

## C.2. Evaluation Benchmarks

**VQAv2 (Goyal et al., 2017)** is a large-scale benchmark for open-ended visual question answering on natural images. The balanced annotation design reduces reliance on language shortcuts and encourages answers to be grounded in the visual content.It consists of over 1.1 million questions derived from more than 200,000 images from the MS COCO dataset (Lin et al., 2014), with each question featuring 10 human-annotated answers.

**GQA (Hudson & Manning, 2019)** targets compositional visual reasoning through scene-graph annotations of real-world images, evaluating whether a model can reason about objects, attributes, and their relations in a structured and multi-step manner. The dataset comprises roughly 22 million questions associated with over 113,000 images from Visual Genome, heavily utilizing scene graphs for structured question generation.

**ScienceQA (Lu et al., 2022)** serves as a multimodal multiple-choice benchmark composed of science problems paired with

**Algorithm 2** TokenNMS: Query-Aware Semantic Filtering

**Input:** Query-agnostic spatial pruned visual token set $\mathcal{V}^{(1)}$, preserved visual saliency scores $\{s_j\}$, text query tokens $Q$, pre-projection visual features $E = \{e_j\}$, weight parameter $\alpha$, threshold $\tau_2$, token budget $K_2$
**Output:** Pruned and query-grounded visual token set $\mathcal{V}^{(2)}$
Initialize final token set $\mathcal{V}^{(2)} \leftarrow \emptyset$
For each token $v_j \in \mathcal{V}^{(1)}$ do
    Compute semantic alignment score $r_j = S_{\text{sem}}(v_j, Q)$
End For
Normalize $\{r_j\}$ to obtain $\tilde{r}_j \in [0, 1]$ // Min-max normalization
For each token $v_j \in \mathcal{V}^{(1)}$ do
    Compute joint importance $S_{\text{joint}}(v_j) = \alpha \cdot s_j + (1 - \alpha) \cdot \tilde{r}_j$
End For
Sort $\mathcal{V}^{(1)}$ in descending order of $S_{\text{joint}}$ to obtain permutation $\pi_2$
For $j = 1$ to $|\mathcal{V}^{(1)}|$ do
    $v_{\text{cand}} \leftarrow v_{\pi_2(j)}$
    If $\max_{u \in \mathcal{V}^{(2)}} \text{sim}_{\text{vis}}(e_{\text{cand}}, e_u) \leq \tau_2$ then
        $\mathcal{V}^{(2)} \leftarrow \mathcal{V}^{(2)} \cup \{v_{\text{cand}}\}$ // Decoupled spatial NMS
    End If
    If $|\mathcal{V}^{(2)}| == K_2$ then
        break // Reach final target budget
    End If
End For
Return $\mathcal{V}^{(2)}$

textual context and visual content. Following standard practice in VLM evaluation, we report results on the image subset to assess visually grounded scientific reasoning. The full dataset encompasses over 21,000 questions sourced from elementary and high school science curricula, categorized across diverse topics such as biology, physics, and chemistry.

**TextVQA (Singh et al., 2019)** measures the ability to answer questions that depend on reading text embedded in natural scenes. Strong performance requires both accurate text perception and reasoning with the surrounding visual context. The benchmark is built upon around 28,000 images from the Open Images dataset, containing over 45,000 questions that explicitly require reading and interpreting scene text.

**POPE (Li et al., 2023)** is designed to evaluate object hallucination in multimodal models through object-presence queries, and is widely used to test whether model responses remain faithful to the actual visual content. The evaluation framework formulates binary yes-or-no questions across three distinct sampling settings—random, popular, and adversarial—using images from MS COCO and other generated sources.

**MME (Fu et al., 2023)** offers a broad multimodal evaluation suite that spans both visual perception and higher-level reasoning tasks, covering abilities such as text reading, recognition, counting, spatial interpretation, and commonsense reasoning. It is structured into 14 distinct subtasks evaluated through standardized binary questions to measure both perception and cognition capabilities systematically.

**MMBench (Liu et al., 2024c)** provides a large-scale multiple-choice framework for the standardized evaluation of multimodal models, incorporating diverse question types to assess perception and reasoning ability in a unified setting. The benchmark contains approximately 3,000 multiple-choice questions meticulously categorized into 20 fine-grained ability dimensions structured within a three-level hierarchical taxonomy.

**MMVet (Yu et al., 2023)** challenges models to combine multiple vision-language skills within a single open-ended response, making it particularly useful for evaluating holistic multimodal reasoning beyond isolated recognition tasks. It is composed of 218 open-ended questions paired with 200 images, specifically designed to test the integration of six core capabilities including recognition, spatial awareness, and OCR.

**ChartQA (Masry et al., 2022)** focuses on question-answering over charts and plots, requiring an understanding of both the visual structure and the underlying numeric information for multi-step or arithmetic reasoning. The dataset includes about 32,000 charts paired with over 31,000 questions, blending machine-generated queries for structural extraction and

*Table 5.* Ablation results on the effectiveness of our proposed framework using LLaVA-1.5-7B across various benchmarks.

| Stage 1 | Stage 2 | Method | *Retain 128 Tokens* | | | |
|---|---|---|---|---|---|---|
| | | | GQA | VQA$^{\text{Text}}$ | POPE | MME |
| ✓ | ✗ | Top-$K$ | 55.2 | 48.5 | 82.1 | 1345.5 |
| ✓ | ✗ | Ours | 56.8 | 50.2 | 85.5 | 1390.2 |
| ✗ | ✓ | Top-$K$ | 56.5 | 54.8 | 83.2 | 1378.0 |
| ✗ | ✓ | Ours | 57.5 | 55.4 | 84.1 | 1405.5 |
| ✓ | ✓ | Top-$K$ | 58.3 | 55.8 | 84.5 | 1422.7 |
| ✓ | ✓ | **Ours** | **60.0** | **56.6** | **87.3** | **1461.3** |

*Table 6.* Ablation results on decoupled NMS and joint ranking in query-aware semantic filtering.

| Distance Space | | Ranking Criteria | | *Retain 128 Tokens* | | | |
|---|---|---|---|---|---|---|---|
| Semantic | Visual | Relevance $r_j$ | Blended $S_{\text{joint}}$ | GQA | VQA$^{\text{Text}}$ | POPE | MME |
| ✓ | ✗ | ✓ | ✗ | 53.5 | 52.1 | 80.2 | 1320.5 |
| ✓ | ✗ | ✗ | ✓ | 54.8 | 53.5 | 81.5 | 1350.0 |
| ✗ | ✓ | ✓ | ✗ | 57.2 | 55.1 | 85.8 | 1410.2 |
| ✗ | ✓ | ✗ | ✓ | **60.0** | **56.6** | **87.3** | **1461.3** |

human-written queries for complex reasoning.

**AI2D (Kembhavi et al., 2016)** evaluates diagram understanding based on science diagrams from educational materials, testing whether a model can interpret diagram components, their relations, and the meaning conveyed by structured visual layouts. It consists of nearly 5,000 educational science diagrams annotated with rich structural representations, accompanied by over 15,000 multiple-choice questions.

**OCRBench (Liu et al., 2024d)** aggregates diverse text-related tasks across scene and document images to evaluate text perception, text-centric reasoning, and robustness in visually rich language understanding. By integrating 29 distinct OCR-centric datasets, it provides 1,000 carefully curated question-answer pairs that evaluate five specific text-related abilities.

**HallusionBench (Guan et al., 2024)** diagnoses multimodal hallucination under visually confusing or linguistically misleading conditions, examining whether predictions remain properly grounded in the image rather than drifting towards unsupported responses. The benchmark comprises 346 images and 1,092 question-answer pairs, purposefully pairing original images with manipulated versions and control questions to strictly evaluate visual and knowledge illusions.

## D. Ablation Studies

### D.1. Effectiveness of Two-Stage Framework

We conducted an ablation experiment on the effectiveness of TokenNMS to demonstrate the necessity of both pruning stages and the superiority of our suppression method over the conventional Top-$K$ approach under a 128-token budget. As shown in Table 5, relying solely on stage 1 (*i.e.*, query-agnostic spatial pruning) yields lower performance on text-dependent tasks such as TextVQA (Singh et al., 2019), although our method reduces hallucinations compared to Top-$K$ validated by a higher POPE (Li et al., 2023) score. Conversely, leveraging only stage 2 (*i.e.*, query-aware semantic filtering) improves text alignment, but lacks the spatial diversity required for complex spatial reasoning, resulting in suboptimal GQA (Hudson & Manning, 2019) scores. Integrating both stages with our suppression method achieves optimal overall performance, demonstrating that systematically decoupling spatial and semantic pruning is crucial to preserving both global context and fine-grained visual details.

Decoupled NMS and Joint Ranking We ablated the core components of our query-aware semantic filtering stage in Table 6. Simply computing the NMS distance metric in the post-projection semantic space suffers from drastic performance degradation, as shown in the first and second rows. This occurs because visually distinct objects sharing similar semantic concepts are erroneously suppressed. By computing the distance metric in the pre-projection visual space, our TokenNMS preserves fine-grained spatial diversity, significantly improving the performance. Furthermore, relying solely on text relevance $r_j$ for token ranking limits the capacity of the model to retain crucial background context, as shown in the third row. Employing $S_{\text{joint}}$ optimally balances visual saliency and semantic alignment, producing optimal reasoning capabilities across all benchmarks.

### D.2. Hyperparameter Sensitivity Analysis

To validate the robustness and optimal configuration of our framework, we conducted hyperparameter sensitivity analysis on the GQA (Hudson & Manning, 2019) and POPE (Li et al., 2023) benchmarks. Figure 4 (a) illustrates the GQA performance across different combinations of the spatial threshold $\tau_1$ and the decoupled threshold $\tau_2$. The results exhibit minimal variance,

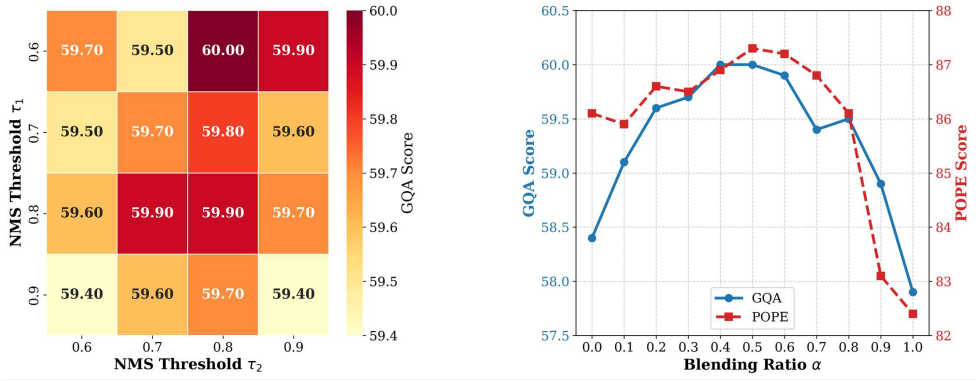


*Figure 4.* Hyperparameter sensitivity analysis. (a) GQA score variations across different NMS thresholds $\tau_1$ and $\tau_2$. The minimal variance indicates high robustness, with the optimal setting at $\tau_1 = 0.6$ and $\tau_2 = 0.8$. (b) Sensitivity of GQA and POPE scores to the blending ratio $\alpha$. The inverted U-shape demonstrates that the optimal balance between inherent visual structure and query-driven semantics is achieved at $\alpha = 0.5$.

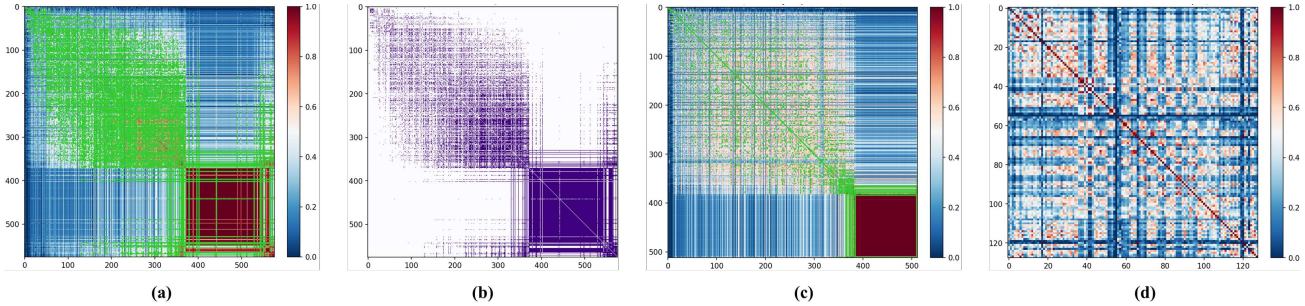


*Figure 5.* Visualization of token similarity across the two-stage pruning process. (a) Initial similarity matrix sorted by visual saliency. (b) Suppression graph during stage 1 spatial NMS. (c) Intermediate similarity matrix re-sorted by joint importance $S_{\text{joint}}$. (d) Final selected tokens after stage 2, preserving spatial diversity.

demonstrating that our decoupled NMS strategy is highly robust to threshold choices. Notably, the peak performance is achieved at $\tau_1 = 0.6$ and $\tau_2 = 0.8$, which empirically validates our design choice of employing conservative spatial pruning followed by relaxed semantic filtering. Furthermore, Figure 4 (b) visualizes the impact of the blending ratio $\alpha$ in the joint importance scoring. Relying solely on query-driven semantic relevance (*i.e.*, $\alpha = 0.0$) or purely on visual saliency (*i.e.*, $\alpha = 1.0$) leads to suboptimal performance, as it sacrifices either spatial structural diversity or task-specific focus. The results validates that equally prioritizing inherent visual structure and semantic relevance (*i.e.*, $\alpha = 0.5$) is crucial for extracting the most informative visual tokens.

### D.3. NMS Suppression Analysis

To demonstrate how our framework mitigates token redundancy, we visualized the token similarity matrices throughout the pruning process. Figure 5 (a) shows the initial similarity matrix sorted by visual saliency, where extensive regions of feature homogeneity indicate spatial redundancy. In the first stage, our spatial NMS effectively mitigates these dense clusters, as illustrated by the suppression graph in Figure 5 (b). After spatial pruning, the remaining tokens are re-sorted based on their joint importance score $S_{\text{joint}}$. The intermediate similarity matrix in Figure 5 (c) demonstrates a reduction in structural bias, although semantic overlaps still remain. Finally, Figure 5 (d) visualizes the tokens retained after the query-aware semantic filtering. The resulting matrix exhibits minimal semantic overlap where self-similarity is maximized along the diagonal, while the off-diagonal regions remain predominantly distinct. This demonstrates that TokenNMS successfully discards redundant information, yielding a highly sparse, mutually exclusive, and structurally diverse set of visual tokens.

*Table 7.* Ablation results on initial token budget $K_1$ using LLaVA-1.5-7B across various benchmarks.

| $K_1$ | *Retain 128 Tokens* | | | |
|---|---|---|---|---|
| | **GQA** | **VQA$^{\text{Text}}$** | **POPE** | **MME** |
| 128 | 56.8 | 50.2 | 85.5 | 1390.2 |
| 256 | 57.7 | 55.6 | 86.2 | 1445.7 |
| 512 | **60.0** | **56.6** | **87.3** | **1461.3** |
| 576 | 57.5 | 55.4 | 84.1 | 1405.5 |

### D.4. Analysis of the Initial Token Budget $K_1$

We conducted an additional experiment on LLaVA-1.5-7B under a 128 token budget using the GQA, TextVQA, POPE and MME benchmarks to demonstrate the impact of the initial token budget $K_1$ on reasoning capabilities. As shown in Table 7, setting $K_1$ to 512 achieves the optimal performance across all evaluated benchmarks. A smaller budget degrades performance because it discards necessary background context based solely on visual scores before considering the text query. Conversely, using all tokens (*i.e.*, $K_1$ =576) also reduces performance, as redundant visual noise interferes with the subsequent semantic filtering stage. Setting $K_1$ to 512, which is approximately 88% of visual tokens, provides the best trade-off by effectively removing obvious spatial redundancy while preserving enough diverse candidate tokens for the query-aware stage.

### D.5. Visualizations on Joint Importance Scores

In Fig. 6, we provide visualization of joint importance scores $S_{joint}$ on the POPE benchmark across varying text queries using LLaVA-1.5-7B. The heatmaps demonstrate that the proposed joint scoring mechanism dynamically adapts token relevance based on the specific semantic instruction. As shown across the examples, the high-scoring regions, indicated in red, accurately localize the target objects regardless of their scale or position. The metric effectively captures fine-grained details, such as a toothbrush or backpack, as well as broader contextual elements, like a couch or car, corresponding to the exact text prompt. Simultaneously, visually salient but query-irrelevant regions are assigned low scores, indicated in blue. This confirms that the joint importance metric successfully alignes raw visual features with text semantics, ensuring that the subsequent pruning stage isolates only task-critical visual tokens while suppressing uninformative background noise.

## E. Limitations

While TokenNMS demonstrates robust performance across diverse benchmarks, it presents certain limitations. First, as a training-free framework, its efficacy inherently depends on the quality of attention maps and pre-projection features of the base Vision Transformer (ViT). If the underlying ViT fails to assign sufficient initial saliency to a tiny but task-critical object, the query-agnostic stage may inadvertently discard it before text-visual alignment occurs. Second, the framework relies on predefined parameters such as $\tau_1$, $\tau_2$, $K_1$ and $\alpha$. Although empirical analysis indicates general robustness to these parameters, applying fixed configurations across images with drastically different structural complexities may lead to suboptimal token retention. Finally, in scenarios with extreme feature homogeneity , the strict Non-Maximum Suppression (NMS) constraint can over-suppress tokens, causing a token deficit. While this is resolved by deterministically padding the sequence with the highest-scoring unselected candidates, this fallback mechanism risks reintroducing localized redundancy, slightly diminishing the spatial diversity enforced by the NMS.

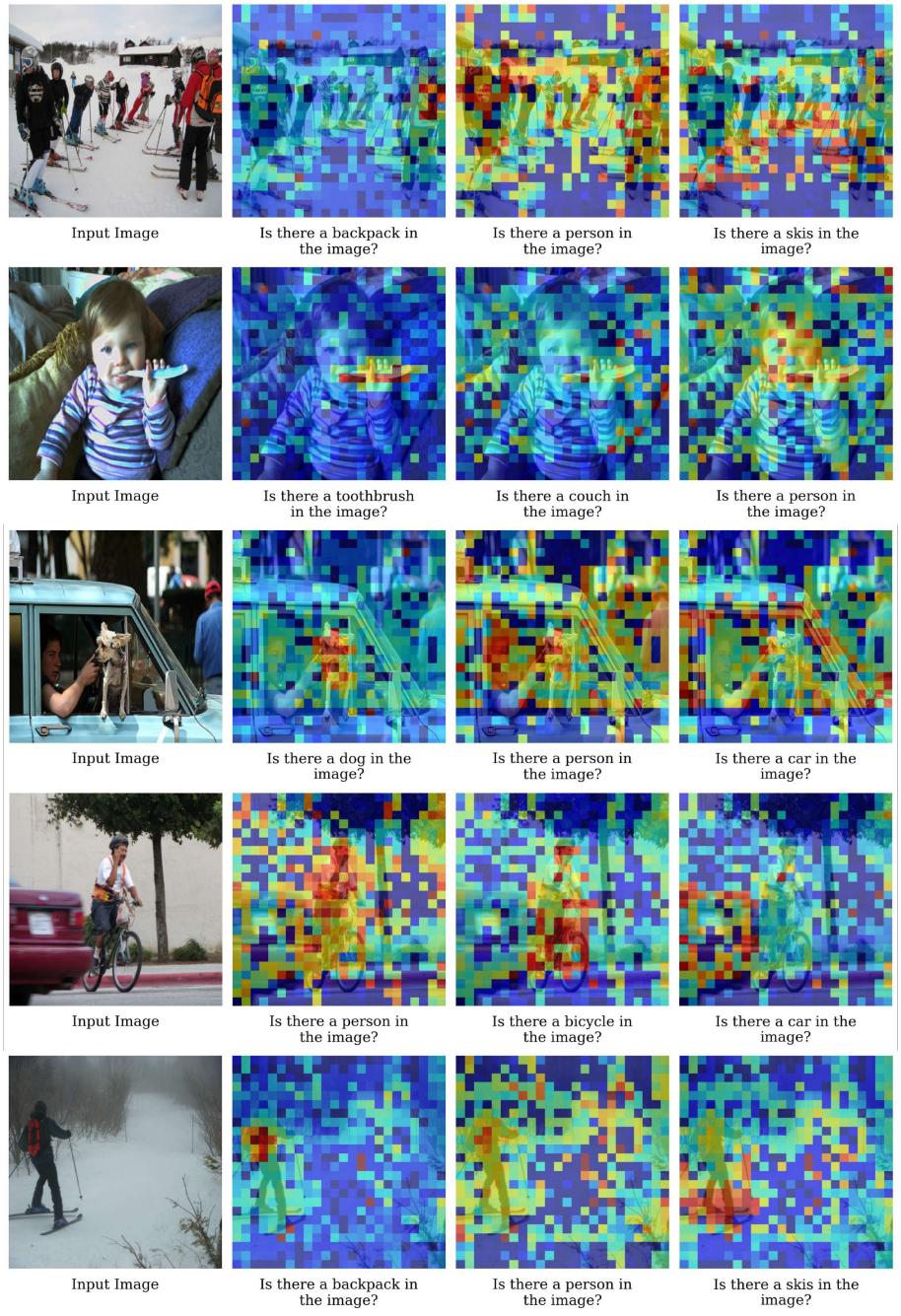


*Figure 6.* Visualizations on joint importance scores on POPE benchmark. Red refers to high score, while blue indicates low.