VLM-Pruner: Buffering for Spatial Sparsity in an Efficient VLM Centrifugal Token Pruning Paradigm

Supplementary Material

6. More Comparative Experiment

6.1. More Baselines

we select CDPruner [45], BTP [7], and SAINT [16] as baselines for more theoretical and empirical comparisons with VLMPruner. To further demonstrate VLMPruner's practical timeliness and applicability, we adapt it to the Qwen3-VL-4B vision-language model and conduct additional comparisons. As shown in Table 7, VLMPruner exhibits a clear advantage while retaining only 11.1% of visual tokens. Compared to other methods, **(1) CDPruner [45]** focuses on **importance-driven** pruning by selecting tokens based on joint relevance to both text and visual tokens, but it may miss critical local details. **(2) BTP [7]**, while **redundancy-reduction**, disrupts fine-grained details early in the process with shallow-layer pruning, rendering later importance-driven pruning less effective. **(3) SAINT [16]**, relying purely on **redundancy-reduction** pruning, may lose important local features due to its lack of spatial consideration. VLMPruner balances diversity and detail completeness by prioritizing local information through centrifugal expansion. This selects more relevant tokens while preserving fine-grained details, resulting in a more balanced selection process that excels in fine-grained visual tasks such as OCR and outperforming the second-best SAINT method by 2.15%.

Method	GQA	MMB	MME	POPE	SQA	VQA^{Text}	OCRBench	OK-VQA	Avg.
Qwen3-VL-4B	62.97	84.02	2327	90.10	92.56	77.94	835	47.42	100.0%
Retain About 120 Tokens (↓ 88.9%)									
SAINT	58.64	**81.11**	2025	83.13	**85.87**	63.75	566	41.59	87.38%
BTP	42.03	56.22	1424	61.00	76.95	55.45	431	13.27	62.06%
CDPruner	57.24	78.08	2042	86.59	86.12	63.39	486	35.18	84.31%
VLM-Pruner (Ours)	**59.72**	78.48	**2129**	**88.51**	85.23	**68.37**	**573**	**42.60**	**89.53%**

Table 7. Comparative experiments on image understanding are performed on Qwen3-VL-4B-Instruct.

6.2. More Visualizations

VLMPruner's centrifugal expansion design is based on the assumption that visual tokens exhibit spatial locality, prioritizing the retention of important information within local neighborhoods. By balancing diversity and detail completeness through the combination of redundancy and spatial sparsity, VLMPruner ensures that nearby tokens are more likely to be selected, avoiding detail loss from dispersed coverage. As shown in Fig. 5, this reduces edge tokens' number and improves detail completeness, leading significant improvements in tasks like fine-grained recognition. Firstly, because edge tokens exhibit extremely low similarity, DART and DivPrune select them far more frequently than VLMPruner. Moreover, VLMPruner preserves fine-grained details more effectively than DART and DivPrune, such as regions highlighted by the red circles in Fig. 5.

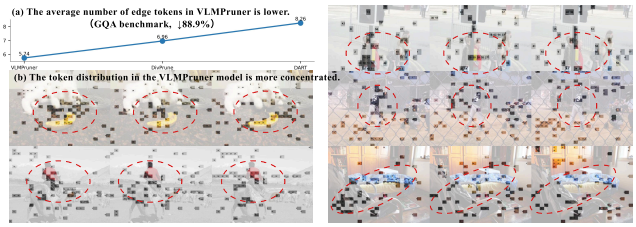

Figure 5. More visualizations of the actual pruning effects between baselines and VLM-Pruner. From left to right are VLM-Pruner, DivPrune, and DART. (a) The average number of edge tokens in VLM-Pruner is lower. (b) The token distribution in the VLM-Pruner model is more concentrated.

7. More Ablation Study

To further validate the effectiveness of token pruning by VLM-Pruner and the reasonableness of the hyperparameter settings, we present several additional experimental results on LLaVA-1.5-7B with the pruning rate of 88.9%, focusing on (a) number of pivots κ, (b) top-q highest variance channels, (c) threshold $\tau^{(0)}$ of token selection, (d) token batch size B, (e) aggregation weight β, and (f) pruning layer i.

7.1. Ablation Study on the Number of Pivots κ

In the experiment on the number of pivots as shown in Table 8, we observe that the performance varied with different values of κ. The optimal performance was achieved with $\kappa = 4$, resulting in an average performance of 95.30%. This suggests that four pivots strike the best balance, providing sufficient coverage of the target regions without overfitting to a specific area. Fewer pivots (e.g., $\kappa = 1$ or $\kappa = 2$) fail to capture enough semantic diversity, while too many pivots (e.g., $\kappa = 7$ or $\kappa = 8$) tend to overfit and capture redundant tokens from less important regions.

7.2. Ablation Study on the Top-q Highest Variance Channels

For the experiment on selecting the top-q highest variance channels as shown in Table 9, we find that setting $q = 256$ achieved the best performance, with an average score of

Method	MME	POPE	VQA^{T}	OCRBench	SEED^{I}	OK-VQA	Avg.
LLaVA-1.5-7B	*Upper Bound, 576 Tokens* **(100%)**						
Vanilla	1862	85.9	58.17	297	66.18	57.98	100.0%
κ	*Retain 64 Tokens* (↓ **88.9%**)						
1	1737	81.6	55.93	271	61.83	56.29	94.36%
2	1729	81.9	55.66	271	62.33	56.52	94.46%
3	1725	**82.4**	55.80	269	**62.36**	56.33	94.41%
4	**1752**	82.2	56.05	279	62.24	56.63	**95.30%**
5	1734	81.6	55.97	273	62.21	56.72	94.68%
6	1732	82.0	**56.11**	**281**	62.17	**56.81**	95.25%
7	1722	81.5	55.88	274	62.11	56.67	94.55%
8	1707	82.0	56.03	273	62.13	56.33	94.40%

Table 8. Ablation study on the number of pivots κ.

95.30%. This configuration retains enough channels to preserve essential visual information while minimizing redundancy. Larger values of q result in diminishing returns, as they introduce more redundant information and computational complexity, which doesn't contribute significantly to performance improvement. On the other hand, smaller values of q lead to the loss of critical visual features, reducing the overall model performance. Therefore, $q = 256$ is the optimal setting, offering a balanced trade-off between computational efficiency and model performance.

Method	MME	POPE	VQA^{T}	OCRBench	SEED^{I}	OK-VQA	Avg.
LLaVA-1.5-7B	*Upper Bound, 576 Tokens* **(100%)**						
Vanilla	1862	85.9	58.17	297	66.18	57.98	100.0%
Top-q	*Retain 64 Tokens* (↓ **88.9%**)						
128	**1765**	81.8	**56.25**	277	62.02	56.70	95.25%
256	1752	**82.2**	56.05	**279**	62.24	56.63	**95.30%**
512	1738	81.3	55.87	270	62.21	56.59	94.42%
1024	1743	81.1	55.51	268	**62.26**	**56.71**	94.26%
4096 (raw)	1752	81.9	55.97	**279**	62.18	56.41	95.14%

Table 9. Ablation study on the Top-q highest variance channels.

7.3. Ablation Study on the Threshold $\tau^{(0)}$ of Token Selection

In the experiment on the token selection threshold $\tau^{(0)}$ as shown in Table 10, we observe that setting $\tau^{(0)} = 0.8$ resulted in the best overall performance of 95.30%. This threshold ensures that the model prioritizes selecting tokens that are spatially close to previously selected ones, effectively avoiding the early inclusion of distant or irrelevant tokens. If the threshold is set too low, distant tokens are selected too early, causing the token set to become scattered and less effective in representing fine-grained details. Conversely, setting the threshold too high causes an overemphasis on spatial distance, resulting in a more compact token selection that introduces increased redundancy. Therefore, a threshold of 0.8 strikes a balance between selecting non-redundant tokens and preserving finer-grained details.

Method	MME	POPE	VQA^{T}	OCRBench	SEED^{I}	OK-VQA	Avg.
LLaVA-1.5-7B	*Upper Bound, 576 Tokens* **(100%)**						
Vanilla	1862	85.9	58.17	297	66.18	57.98	100.0%
$\tau^{(0)}$	*Retain 64 Tokens* (↓ **88.9%**)						
0.6	**1753**	81.9	55.00	276	61.82	56.48	94.64%
0.7	1704	**82.5**	55.68	276	61.93	**56.71**	94.60%
0.8	1752	82.2	**56.05**	**279**	**62.24**	56.63	**95.30%**
0.9	1719	78.3	54.71	269	61.29	56.39	92.99%
1.0	1725	73.9	53.66	265	59.87	55.31	91.00%

Table 10. Ablation study on the threshold $\tau^{(0)}$ of token selection.

7.4. Ablation Study on the Token Batch Size B

In the experiment on token batch size B as shown in Table 11, the results show that a batch size of $B = 16$ provides the best performance and fastest inference speed, with an average score of 95.30% and a total time of 91 minutes. Larger batch sizes, such as $B = 32$, may reduce inference time, but they may lead to the failure of the BSS criterion due to fewer iterations. Smaller batch sizes result in longer inference times, even slower than the initial LLaVA-1.5-7B at 131 minutes. The token batch size of $B = 16$ offers an optimal trade-off, minimizing latency while maintaining the quality of the selected tokens.

Method	MME	POPE	VQA^{T}	OCRBench	SEED^{I}	OK-VQA	Avg.
LLaVA-1.5-7B	*Upper Bound, 576 Tokens* **(100%)**						
Vanilla	131	1862	85.9	58.17	297	66.18	57.98	100.0%
B	Total Time (Min.)	*Retain 64 Tokens* (↓ **88.9%**)						
1	160	**1755**	**83.6**	56.06	275	**62.95**	56.88	95.62%
2	124	1742	83.5	55.96	**279**	62.73	**56.93**	**95.64%**
4	106	1730	83.3	56.11	278	62.57	56.70	95.38%
8	96	1726	83.5	**56.26**	274	62.42	56.82	95.20%
16	**91**	1752	82.2	56.05	**279**	62.24	56.63	95.30%

Table 11. Ablation study on the token batch size B.

7.5. Ablation Study on the Aggregation Weight β

In the aggregation weight β experiment, as shown in Table 12, we focus on minimizing the influence of Stage 3 to emphasize the importance of Stage 1 and 2, which manage buffering for spatial sparsity and token selection, while ensuring the preservation of fine-grained details. The results show that $\beta = 0.3$ yields an average score of 95.30%, providing a moderate improvement over the 95.07% achieved without Stage 3. It also enables effective aggregation of discarded token information without overemphasizing it. This setting further preserves fine-grained details, as demonstrated by the best score of 279 achieved on OCRBench.

To further emphasize the core innovation of VLM-Pruner, the BSS criterion, we validate the model performance on Qwen-VL-7B after removing Stage 3, as shown in Table 13. The results show that even without Stage 3, our method still outperforms existing baselines by an absolute

Method	MME	POPE	VQA^{T}	OCRBench	SEED^{I}	OK-VQA	Avg.
LLaVA-1.5-7B	*Upper Bound, 576 Tokens* **(100%)**						
Vanilla	1862	85.9	58.17	297	66.18	57.98	100.0%
β	*Retain 64 Tokens* (↓ **88.9%**)						
0.2	1737	82.1	55.86	278	62.26	56.41	94.98%
0.3	1752	82.2	56.05	**279**	62.24	56.63	95.30%
0.4	1754	**82.4**	56.26	277	62.28	56.72	95.34%
0.5	1773	**82.4**	**56.44**	278	**62.36**	56.80	95.66%
0.6	1790	82.0	56.16	278	**62.36**	**56.87**	**95.68%**
0.7	1799	81.5	56.23	277	62.33	56.80	95.60%
0.8	**1808**	81.5	56.07	275	62.14	56.81	95.47%
w/o stage 3	1795	81.2	55.72	275	61.91	56.56	95.07%

Table 12. Ablation study on the aggregation weight β.

margin of approximately 5%. This highlights the effectiveness of the proposed centrifugal token pruning paradigm.

Method	MME	POPE	VQA^{T}	OCRBench	SEED^{I}	OK-VQA	Avg.
Qwen2-VL-7B	*Upper Bound* **(100%)**						
Vanilla	2321	88.55	82.73	796	76.65	50.21	100.0%
Stage 3	*Retain 11.1% Tokens* (↓ **88.9%**)						
w/o Stage 3	**2197**	86.4	74.89	561	70.69	48.20	90.25%
FastV	2174	81.6	72.86	454	64.79	46.59	84.72%
DART	2087	81.6	65.10	481	65.02	46.59	83.14%
DivPrune	2059	86.5	71.65	472	70.75	47.29	86.47%
Ours	2158	**87.4**	**76.15**	**581**	**72.14**	**48.35**	**91.20%**

Table 13. Additional structural decomposition analysis.

7.6. Ablation Study on the Pruning Layer i

In the layer-wise pruning experiment summarized in Table 14, pruning at layer 2 achieves the best balance and an average performance of 95.30% by successfully preserving fine-grained visual details while reducing task-irrelevant redundancy. Unlike DivPrune, VLM-Pruner cannot prune layer 0 because it relies on token $keys$ from the previous layer to perform coarse semantic abstraction. Pruning at layer 1 leads to notable performance degradation, likely because the text and visual tokens have not yet fully interacted, which hampers the model's ability to focus on prompt-relevant regions and results in suboptimal token selection. Conversely, pruning at later layers such as layer 3 or 4 may remove more less important details, and this also comes at the cost of lower model performance and increased computational overhead in FLOPs.

Method	MME	POPE	VQA^{T}	OCRBench	SEED^{I}	OK-VQA	Avg.
LLaVA-1.5-7B	*Upper Bound, 576 Tokens* **(100%)**						
Vanilla	1862	85.9	58.17	297	66.18	57.98	100.0%
layer i	*Retain 64 Tokens* (↓ **88.9%**)						
1	1689	82.0	55.68	267	61.54	56.30	93.65%
2	1752	82.2	**56.05**	**279**	62.24	**56.63**	**95.30%**
3	1725	82.9	55.85	274	**62.39**	56.53	94.86%
4	**1759**	**83.3**	55.70	270	62.18	56.48	94.91%

Table 14. Ablation study on the pruning layer i.

The choice of the pruning layer is based on the following reasons: (1) According to the Align-KD [11], feature similarity between layers after the second layer exceeds 0.9, meaning the second layer already contains rich contextual information. Stage 3 compensates for unselected token information, minimizing information loss. (2) Pruning is performed on just one layer, resulting in low computational cost and time consumption. (3) It is easy and fast to adapt to different VLMs with default parameters. In contrast, methods like PDrop [40], BTP [7], and SAINT [16] require manual tuning of pruning layers and thresholds, making adaptation more complex.

7.7. Ablation Study on other base VLMs

To further validate the "plug-and-play" claim of our method, we conduct additional ablation experiments on Qwen3-VL-4B. We select the most critical stages, Stage 1 and Stage 2, for partial benchmarking experiments. As shown in Table 15, the results confirm the strong advantages of the default parameters presented in the paper.

Benchmark	Qwen3-VL-4B	Query Value	$\tau = 0.6$	0.7	0.9	$\lambda = 0.3$	0.7	0.9	Ours
Retain About 120 Tokens (↓ **88.9%**)									
MME	2327	2065	2055	2089	2096	2106	2074	2095	2096	**2129**
SQA	92.56	84.53	85.12	84.23	84.58	84.53	85.18	84.18	84.73	**85.23**
OCRBench	835	571	551	502	550	544	559	557	552	**573**

Table 15. Ablation Study on image understanding are performed on Qwen3-VL-4B-Instruct.

8. Details for Benchmarks

We select a broad range of commonly used tasks and benchmarks. Specifically, the evaluation includes 9 image-language benchmarks, namely GQA [15], MMB [25], MME [5], POPE [20], SQA [27], TextVQA [33], OCR-Bench [26], SEEDBench [18], and OKVQA [29], as well as 4 video-language benchmarks, namely EgoSchema [28], NExTQA [39], VideoMME [12], and EgoPlan [8].

8.1. Image-Language Benchmarks

GQA GQA is designed to evaluate visual scene understanding and compositional reasoning capabilities of VLMs on real-world imagery.

MMB MMBench provides a hierarchical evaluation framework for VLMs, organizing 20 distinct capability dimensions into three progressive levels (L1 to L3).

MME MME assesses both perceptual and cognitive abilities of VLMs, where perception includes optical character recognition (OCR), coarse- and fine-grained object recognition, and cognition encompasses commonsense reasoning, numerical calculation, text translation, and code reasoning.

POPE POPE quantifies object hallucination in VLMs by reformulating the problem as a binary judgment task on the presence of objects in an image.

SQA ScienceQA measures scientific reasoning proficiency across natural science, social science, and language science domains, supplemented with detailed explanations and lectures.

TextVQA TextVQA evaluates the integration of visual and textual information, requiring models to read and reason about text present in images.

OCRBench OCRBench provides a comprehensive evaluation of text-related capabilities, covering representative tasks such as scene text recognition, document-oriented understanding, key information extraction, and handwritten mathematical expression recognition.

SEEDBench SEEDBench assesses multimodal understanding across text and visual modalities, including scene classification, object detection, and attribute recognition.

OKVQA OKVQA examines the ability of VLMs to leverage external knowledge, as the image content alone is insufficient to answer the questions.

8.2. Video-Language Benchmarks

EgoSchema EgoSchema focuses on long-term video-language understanding, requiring capabilities such as scene comprehension, object state tracking, and extended visual memory.

NExTQA NExTQA challenges VLMs to interpret temporal and causal structures in videos, emphasizing reasoning about action causality, temporal progression, and object interactions.

VideoMME VideoMME is the first benchmark dedicated to holistic video analysis, systematically evaluating model performance across varying video durations (short, medium, long) and input modalities (video frames, subtitles, audio).

EgoPlan EgoPlan assesses VLMs as embodied task planners, involving interactions with hundreds of objects under complex visual settings to evaluate the prediction of feasible actions.