Beyond Text-Visual Attention: Exploiting Visual Cues for Effective Token Pruning in VLMs

ICCV 2025 Poster

Qizhe Zhang[1] Aosong Cheng[1] Ming Lu[1] Renrui Zhang[3] Zhiyong Zhuo[1]
Jiajun Cao[1] Shaobo Guo[2] Qi She[2] Shanghang Zhang[1]

[1]Peking University [2]ByteDance [3]CUHK MMLab

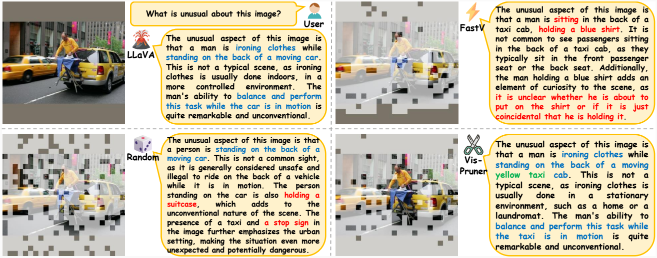
Motivation: The Problems within VLMs

Vision-Language Models (VLMs) are computationally expensive.

- Visual inputs generate far more tokens than text inputs (*e.g.*, LLaVA-NeXT has 2880 tokens).
- This creates a significant bottleneck for inference speed and memory cost.

Existing Solution: Visual Token Pruning for VLMs.

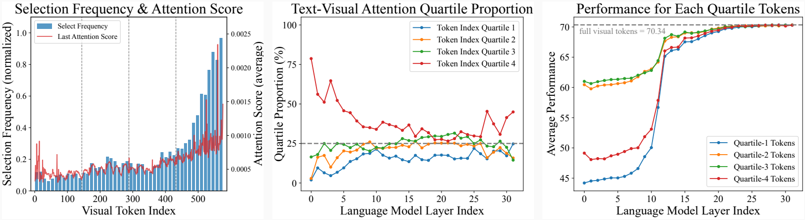
- Previous methods prune redundant visual tokens mainly based on text-visual attentions from the language model.
- **Our Finding:** This is NOT an ideal indicator for pruning.



Phenomenon 1: Text-Visual Attention Shift

Finding

Text-visual attention suffers from a strong **positional bias**. Text tokens pay more attention to visual tokens that are physically closer to them in the sequence (*i.e.*, the bottom of the image).

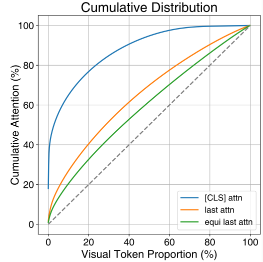


(a) Positional Bias: Attention score and token selection frequency are skewed towards later tokens. **(b) Layer-wise Bias:** The positional bias is strongest in early layers. **(c) Performance Drop:** Tokens with the highest attention (Quartile-4) do not yield the best performance in early layers; central tokens (Quartile-2,3) are more important.

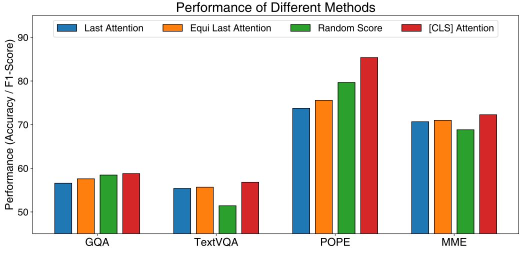
Phenomenon 2: Text-Visual Attention Dispersion

Finding

Even after removing positional bias, text-visual attention is too **dispersed** (high-entropy), making it difficult to distinguish truly important tokens from less important ones.



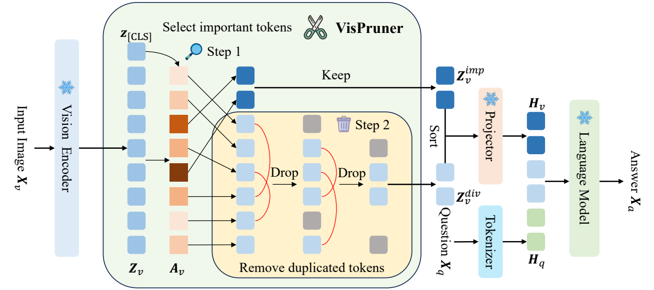
CDF: [CLS] attention is highly concentrated. Text-visual attention in LLM is spread out.



Pruning Performance: Pruning with dispersed last attention struggles and can be worse than random. Pruning with concentrated [CLS] attention is consistently the best.

Our Method: VisPruner

We propose **VisPruner**, a training-free method that exploits visual cues from the visual encoder before the language model.



Step 1: Select Important Tokens

- Use **[CLS] attention** from the visual encoder to identify the most information-rich tokens (e.g., foreground objects).

Step 2: Select Diverse Tokens

- From the remaining tokens, remove redundant ones based on cosine similarity to capture diverse background information.

Main Results: LLaVA-1.5

VisPruner consistently outperforms other methods across various token reduction ratios on 10 benchmarks.

Method	VQA^{V2}	GQA	VizWiz	SQA^{IMG}	VQA^{Text}	POPE	MME	MMB	MMB^{CN}	MMVet	Acc.	Rel.
Upper Bound, All 576 Tokens (**100%**)												
LLaVA-1.5-7B	78.5	62.0	50.0	66.8	58.2	85.9	1510.7	64.3	58.3	31.1	63.1	100.0%
Retain 128 Tokens (↓ **77.8%**)												
FastV	61.8	49.6	51.3	60.2	50.6	59.6	1208.9	56.1	51.4	28.1	52.9	85.4%
SparseVLM	73.8	56.0	51.4	67.1	54.9	80.5	1376.2	60.0	51.1	30.0	59.4	94.4%
VisionZip	75.6	57.6	52.0	68.9	56.8	83.2	1432.4	62.0	56.7	32.6	61.7	98.4%
VisPruner (Ours)	**75.8**	**58.2**	**52.7**	**69.1**	**57.0**	**84.6**	**1461.4**	**62.7**	**57.3**	**33.7**	**62.4**	**99.6%**
Retain 64 Tokens (↓ **88.9%**)												
FastV	55.0	46.1	50.8	51.1	47.8	48.0	1019.6	48.0	42.7	25.8	46.6	75.9%
SparseVLM	68.2	52.7	50.1	62.2	51.8	75.1	1221.1	56.2	46.1	23.3	54.7	86.4%
VisionZip	72.4	55.1	52.9	69.0	55.5	77.0	1365.6	60.1	**55.4**	31.7	59.7	95.6%
VisPruner (Ours)	**72.7**	**55.4**	**53.3**	**69.1**	**55.8**	**80.4**	**1369.9**	**61.3**	55.1	**32.3**	**60.4**	**96.6%**
Retain 32 Tokens (↓ **94.4%**)												
FastV	43.4	41.5	51.7	42.6	42.5	32.5	884.6	37.8	33.2	20.7	39.0	64.1%
SparseVLM	58.6	48.3	51.9	57.3	46.1	67.9	1046.7	51.4	40.6	18.6	49.3	77.9%
VisionZip	67.1	51.8	52.9	68.8	53.1	68.7	1247.4	57.7	50.3	25.5	55.8	89.0%
VisPruner (Ours)	**67.7**	**52.2**	**53.0**	**69.2**	**53.9**	**72.7**	**1271.0**	**58.4**	**52.7**	**28.8**	**57.2**	**91.5%**

With **77.8%** of the visual tokens pruned, VisPruner remarkably maintains **almost all** of the original performance. At an extreme **94.4%** reduction ratio, VisPruner still maintains **91.5%** of the original performance.

Performance on High-Resolution and Video

VisPruner's effectiveness scales to scenarios with higher token counts.

High-Resolution (LLaVA-NeXT)

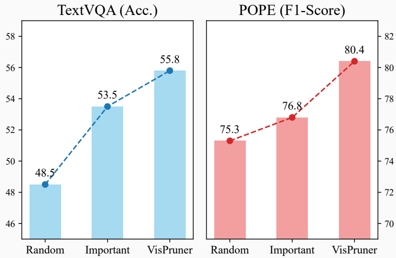
Method	Acc.	Rel.
Upper Bound, All 2880 Tokens (**100**%)		
LLaVA-NeXT-7B	73.1	100.0%
Retain 640 Tokens (↓ **77.8**%)		
FastV	70.9	97.0%
VisPruner (Ours)	**72.1**	**98.6%**
Retain 320 Tokens (↓ **88.9**%)		
FastV	63.9	87.7%
VisPruner (Ours)	**68.1**	**93.3%**
Retain 160 Tokens (↓ **94.4**%)		
FastV	53.8	74.7%
VisPruner (Ours)	**63.1**	**86.7%**

Video QA (Video-LLaVA)

Method	Acc.	Score
Upper Bound, All 2048 Tokens (**100**%)		
Video-LLaVA-7B	49.3	3.32
Retain 455 Tokens (↓ **77.8**%)		
FastV	47.6	3.28
VisPruner (Ours)	**48.4**	**3.31**
Retain 227 Tokens (↓ **88.9**%)		
FastV	45.4	3.24
VisPruner (Ours)	**46.9**	**3.26**
Retain 114 Tokens (↓ **94.4**%)		
FastV	42.4	3.15
VisPruner (Ours)	**44.5**	**3.18**

Ablation and Efficiency

Ablation Study



Both Important and Diverse tokens are crucial. Combining them yields the best performance (**VisPruner**).

Efficiency Analysis

Method	FLOPs (T)	Stroge (MB)	Latency (ms)
Upper Bound, All 2880 Tokens (**100%**)			
LLaVA-NeXT-7B	43.6	1440	313
Retain 640 Tokens (↓ **77.8%**)			
FastV	13.5	380	148
VisPruner (Ours)	**11.5**	**360**	**117**
Retain 160 Tokens (↓ **94.4%**)			
FastV	6.3	95	112
VisPruner (Ours)	**3.8**	**80**	**78**

- At the same ratio, VisPruner is faster and more efficient.
- Pruning before LLM is compatible with optimizations like FlashAttention.

Conclusion

Summary of Contributions:

1. We conduct a thorough investigation into text-visual attention, revealing its flaws (shift & dispersion) as an indicator for pruning.
2. We introduce **VisPruner**, a plug-and-play and training-free method that uses visual cues for more effective and efficient token pruning.
3. We demonstrate through extensive experiments that VisPruner consistently outperforms existing methods across various VLM architectures, modalities (image/video), and reduction ratios.

Thank You!

More Resources:

paper **page** **code**

If you have any questions, please contact: `theia@pku.edu.cn`