Nüwa: Mending the Spatial Integrity Torn by VLM Token Pruning

Yihong Huang[1,2], Fei Ma[1,*], Yihua Shao[3], Jingcai Guo[3], Zitong Yu[4], Laizhong Cui[5], Qi Tian[1,6]

1. Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ)
2. School of Artificial Intelligence, Xidian University ; 3. The Hong Kong Polytechnic University
4. Great Bay University; 5. Shenzhen University ; 6. Huawei

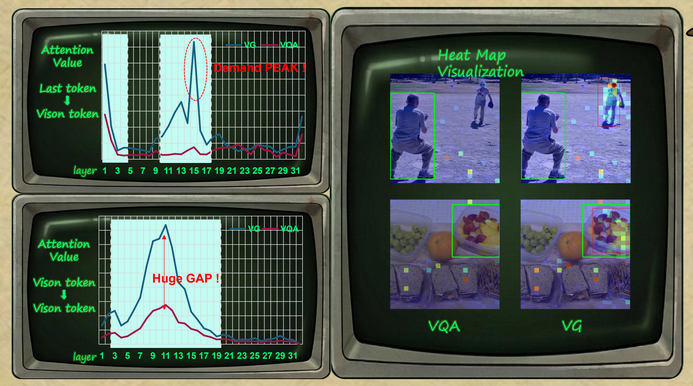
huangyihong@stu.xidian.edu.cn , mafei@gml.ac.cn

I. Revealing the Performance GAP of Token Pruning between VQA and VG Tasks

Performance comparison of various vision token pruning methods on LLAVA1.5-7B.

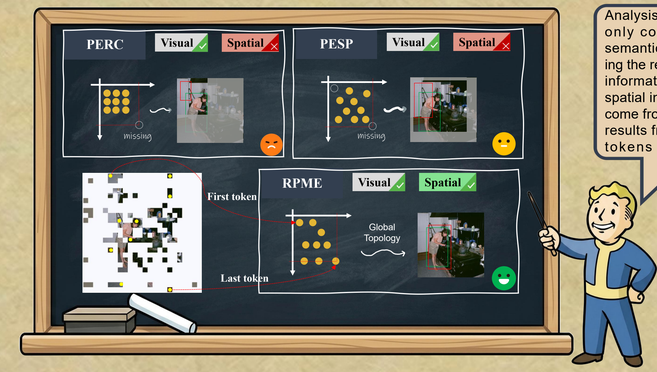
Method	Source	VQA Tasks (Avg Tokens: 64)				Visual Grounding Tasks (Avg Tokens: 64)			
		GQA	MMB	VQAv2	MME	RefCOCO	RefCOCO+(A)	RefCOCO+(B)	RefCOCOg
Vanilla (LLaVA)	CVPR'24	61.9	64.7	78.5	1862	58.30	59.43	38.88	48.50
LLM Single-Layer Pruning									
FastV	ECCV'24	46.1 (74.5%)	48.0 (74.2%)	55.0 (70.1%)	1255 (67.4%)	2.73 (4.7%)	1.17 (2.0%)	1.02 (2.6%)	2.19 (4.5%)
LLM Multi-Layer Pruning									
PDrop	CVPR'25	41.9 (67.7%)	33.3 (51.5%)	57.3 (73.0%)	1092 (58.6%)	–	–	–	–
SparseVLM	ICML'25	53.8 (86.9%)	**60.1** (92.9%)	68.2 (86.9%)	1589 (85.3%)	1.04 (1.8%)	0.96 (1.6%)	1.28 (3.3%)	0.61 (1.3%)
Vision Encoder-Side Pruning									
VisionZip	CVPR'25	55.1 (89.0%)	**60.1** (92.9%)	**72.4** (92.2%)	**1690** (90.8%)	4.04 (6.9%)	3.73 (6.3%)	3.86 (9.9%)	3.38 (7.0%)
PruMerge+	ICCV'25	**55.4** (89.5%)	59.6 (92.1%)	71.3 (90.8%)	1616 (86.8%)	–	–	–	–

WHY: The compression of vision token performs well on VQA tasks but exhibits performance collapse on VG tasks.



Experiments reveal distinct attention patterns across tasks: while sharing common high-attention tokens, the VG task uniquely focuses on target locations. This indicates that compressing semantics is straight-forward, whereas preserving spatial information requires extra attention.

II. Hypothesize, Analyze & Rebuild the spatial integrity



Analysis: Most previous methods only considered compressing semantic information while ignoring the resulting damage to spatial information. We assume that this spatial information does not solely come from a few tokens but rather results from interactions between tokens at different positions.

VisionZip fix from:
PERC ➡ RPME
FastV fix from:
PERC ➡ RPME

Position Reconstruction Experiment on VG and VQA

[Part I] Visual Grounding Tasks *(Significant Improvements)*				
Method	RefCOCO	RefCOCO+(A)	RefCOCO+(B)	RefCOCOg
Vanilla (LLaVA)	58.30	59.43	38.88	48.50
Average 64 Tokens				
VisionZip-fix	11.57 (+7.53)	9.27 (+5.54)	7.57 (+3.71)	8.19 (+4.81)
Fastv-fix	4.52 (+1.79)	3.84 (+2.67)	2.75 (+1.73)	4.17 (+1.98)
Average 128 Tokens				
VisionZip-fix	21.39 (+16.90)	19.96 (+15.90)	13.45 (+8.59)	15.69 (+12.19)
Fastv-fix	13.41 (+3.07)	11.69 (+3.16)	12.29 (+2.46)	12.02 (+3.15)

[Part II] VQA Tasks *(Performance Maintained)*				
Method	GQA	MMB	VQAv2	MME
Vanilla (LLaVA)	61.9	64.7	78.5	1862
Average 64 Tokens				
VisionZip-fix	55.6 (+0.5)	61.8 (+1.7)	70.6 (-1.8)	1700 (+10)
Fastv-fix	46.2 (+0.1)	47.8 (-0.2)	54.1 (-0.9)	1247 (-8)
Average 128 Tokens				
VisionZip-fix	58.5 (+0.9)	63.4 (+1.4)	74.3 (-1.3)	1751 (-10)
Fastv-fix	51.3 (+0.8)	57.7 (+1.6)	60.3 (-1.5)	1494 (+4)

III. The main architectural design of Nüwa project

IV. Main Experiments

Performance comparison of various VQA benchmarks on LLAVA1.5-7B.

Method	Source	GQA	MMB	MMMU	MME	VQAv2	VQAtext	POPE	SQA	SEED	MMVet	avg
Vanilla	CVPR'24	61.9	64.7	36.3	1862	78.5	58.2	85.9	69.5	58.6	31.1	100%
Average Token 192 ↓ 66.7%												
FastV	ECCV'24	52.7	61.2	34.3	1612	67.1	52.5	64.8	67.3	57.1	27.7	89.53%
PDrop	CVPR'25	57.1	63.2	34.1	1766	74.9	56.1	82.3	**70.2**	54.7	30.5	95.87%
SparseVLM	ICML'25	57.6	62.5	33.8	1721	75.6	56.1	83.6	69.1	55.8	31.5	96.11%
VisionZip	CVPR'25	59.3	63.0	**36.6**	1782	**76.8**	57.3	85.3	68.9	56.4	**31.7**	98.26%
Nüwa	-	**60.9**	**64.3**	35.5	**1834**	75.9	**57.4**	**86.4**	68.2	**59.7**	30.5	**98.80%**
Average Token 128 ↓ 77.8%												
FastV	ECCV'24	49.6	56.1	34.9	1490	61.8	50.6	59.6	60.2	55.9	28.1	85.04%
PDrop	CVPR'25	56.0	61.1	34.2	1664	73.5	55.1	82.3	**69.9**	53.3	30.8	94.32%
SparseVLM	ICML'25	56.0	60.0	33.8	1696	73.8	54.9	80.5	67.1	53.4	30.0	93.36%
VisionZip	CVPR'25	57.6	62.0	**37.9**	1761	**75.6**	56.8	83.2	68.9	54.9	**32.6**	97.63%
PruMerge	ICCV'25	57.8	59.6	36.2	1712	74.7	54.3	81.5	67.6	-	30.4	95.06%
Nüwa	-	**60.2**	**63.4**	35.8	**1828**	75.1	**57.0**	**85.5**	67.8	**58.7**	29.8	**97.87%**
Average Token 64 ↓ 88.9%												
FastV	ECCV'24	46.1	48.0	34.0	1256	55.0	47.8	59.6	51.1	51.9	25.8	79.36%
PDrop	CVPR'25	41.9	33.3	26.5	1092	57.3	45.9	55.9	69.2	40.0	24.9	71.56%
SparseVLM	ICML'25	53.8	60.1	35.44	1589	68.2	53.4	77.5	**69.8**	51.1	24.9	89.93%
VisionZip	CVPR'25	55.1	60.1	36.2	1690	72.4	**55.5**	77.0	69.0	52.2	**31.7**	93.99%
PruMerge	ICCV'25	55.4	59.6	35.8	1616	71.3	52.0	75.7	69.5	-	28.0	91.71%
Nüwa	-	**58.3**	**62.0**	**36.4**	**1706**	**72.8**	54.9	**83.0**	67.5	**56.44**	28.2	**94.91%**

Performance comparison of various Refcoco Benchmarks on LLAVA1.5-7B.

Method	Source	Refcoco-test	Refcoco-val	Refcoco+-testA	Refcoco+-testB	Refcoco+-val	Refcocog-test	Refcocog-val	avg
Vanilla	CVPR'24	58.30	56.42	59.43	38.88	46.32	48.50	48.82	100%
Average Tokens 192 ↓ 66.7%									
FEATHER*	ICCV'25	27.7		24.7			27.2		48.38%
Nüwa	-	**47.91**	**46.12**	**43.18**	**31.86**	**37.68**	**37.64**	**37.90**	**79.29%**
Average Tokens 128 ↓ 77.8%									
Fastv	ECCV'24	10.34	10.13	8.53	9.83	8.16	8.87	9.10	18.55%
SparseVLM	ICML'25	6.27	6.17	5.70	4.22	9.85	6.35	6.47	12.84%
VisionZip	CVPR'25	4.49	4.11	4.06	3.86	3.88	3.50	3.48	8.1%
Nüwa		**45.09**	**43.69**	**42.63**	**28.98**	**35.32**	**36.59**	**36.00**	**75.20%**
Average Tokens 64 ↓ 88.9%									
Fastv	ECCV'24	2.73	2.01	1.17	1.02	2.41	2.19	2.01	3.81%
SparseVLM	ICML'25	1.04	1.01	0.96	1.28	0.96	0.61	0.66	1.88%
VisionZip	CVPR'25	4.04	3.81	3.73	3.86	3.50	3.38	3.21	7.28%
Nüwa	-	**29.43**	**28.60**	**28.22**	**17.47**	**22.22**	**21.81**	**21.42**	**47.19%**

Ablation Study on each design. Include Pillar-token selecting, Stage2 Random Pruning and Region Separation

region	pillar token	random S2	GQA	MMB	MME	Refcoco test	Refcoco+ testA	Refcoco+ testB	Refcocog test
✘	✘	✘	58.84	58.18	1791	6.83	6.54	4.50	5.58
✘	✘	✔	57.07	56.43	1736	6.72	6.48	4.38	5.25
✘	✔	✘	59.62	62.98	1807	6.35	6.12	4.65	3.60
✘	✔	✔	58.41	61.71	1771	7.01	6.81	4.42	4.92
✔	✘	✘	57.94	56.68	1742	43.50	39.85	26.10	34.20
✔	✘	✔	57.35	56.10	1724	43.17	38.94	25.58	33.74
✔	✔	✘	**60.18**	**63.40**	**1828**	**45.09**	**42.63**	**28.98**	**36.59**
✔	✔	✔	59.03	62.14	1791	44.30	41.20	27.50	35.80

Ablation Study On cohesion distance of Stage1's Region Cohesion.

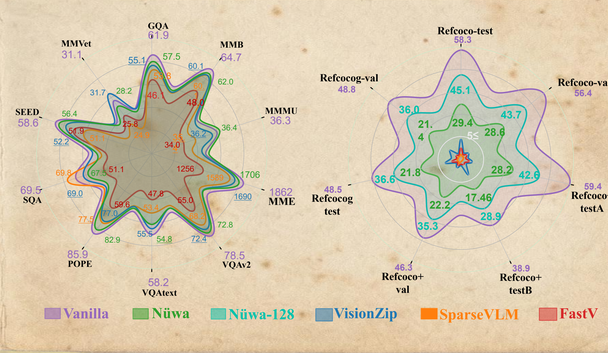
Config	GQA	MMB	MME	Refcoco -test	Refcoco+ -testA	Refcoco+ -testB	Refcocog -test
dist18	0.5784/0	60.2852	1695.37	0.2783	0.2818	0.1655	0.2018
dist148	**0.5853**	61.6838	1704.981	0.2922	**0.2936**	0.1730	**0.2189**
dist280	0.5833	62.0275	1706.869	**0.2943**	0.2822	**0.1747**	0.2181
dist412	0.5826	**62.1134**	**1711.202**	0.2829	0.2705	0.1698	0.2155
dist544	0.5811	62.0275	1702.986	0.2834	0.2637	0.1651	0.2100
dist676	0.5810	61.9416	1698.32	0.2801	0.2590	0.1636	0.2083
dist808	0.5808	61.8557	1706.869	0.2769	0.2607	0.1632	0.2071
dist940	0.5799	61.9416	1705.986	0.2774	0.2572	0.1622	0.2059
dist1058	0.5799	61.7698	1704.486	0.2765	0.2560	0.1630	0.2054

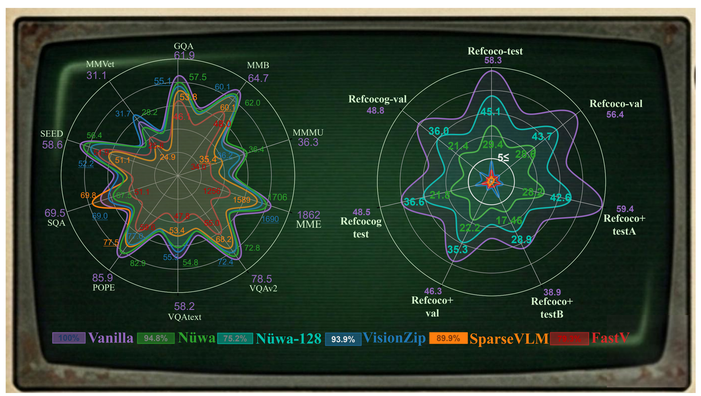
Performance comparison of various VQA benchmarks on Qwen2.5-VL 7B.

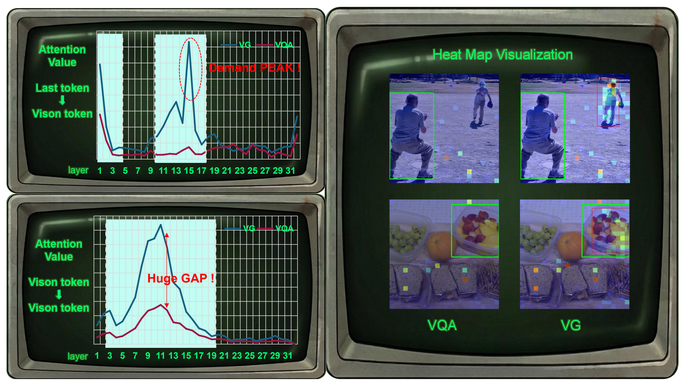
Methods	GQA	POPE	SQAimg	MMB-en	MME	VQA_text	avg
Vanilla	**61.9**	**87.9**	**77.8**	**83.5**	**2347**	**82.2**	**100.0%**
Average Tokens 75%							
Nüwa	60.41	87.52	77.98	83.13	2340	77.35	98.5%
Average Tokens 50%							
Nüwa	59.93	87.46	78.82	83.02	2330	76.03	98.1%
Average Tokens 25%							
Nüwa	58.4	87.06	78.58	82.47	2313	73.81	96.9%

Performance comparison of various Refcoco Benchmarks on Qwen2.5-VL 7B.

Method	Refcoco-testA	Refcoco-testB	Refcoco+-testA	Refcoco+-testB	Refcocog-test	avg
Vanilla	**92.56**	**85.16**	**89.02**	**79.15**	**87.24**	**100%**
Average Tokens 75%						
Nüwa	91.76	84.37	87.98	77.18	86.87	98.8%
Average Tokens 50%						
Nüwa	90.04	82.85	86.74	72.65	85.49	96.4%
Average Tokens 25%						
Nüwa	80.71	72.83	73.57	62.4	73.96	83.8%




GQA
63.9
MMB
64.7
MMMU
36.3
MME
1862
VQAv2
78.5
VQAtext
58.2
POPE
85.9
SQA
69.5
SEED
58.6
MMVet
31.1
Refcoco-test
58.3
Refcoco-val
56.4
59.4
Refcoco+
testA
38.9
Refcoco+
testB
46.3
Refcoco+
val
48.5
Refcocog
test
Refcocog-val
48.8
100% Vanilla
94.6% Nüwa
75.2% Nüwa-128
93.9% VisionZip
89.9% SparseVLM
FastV


Attention
Value
Last token
⬇
Vison token
VG
VQA
Demand PEAK !
layer 1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
Attention
Value
Vison token
⬇
Vison token
Huge GAP !
VG
VQA
layer 1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31

Heat Map Visualization
VQA
VG