Beyond Selection

Spotlight



Token Parameterization for Extreme Visual Token Compression

Rui Zhong, Yu Li, Zheyu Yan ✉, Cheng Zhuo

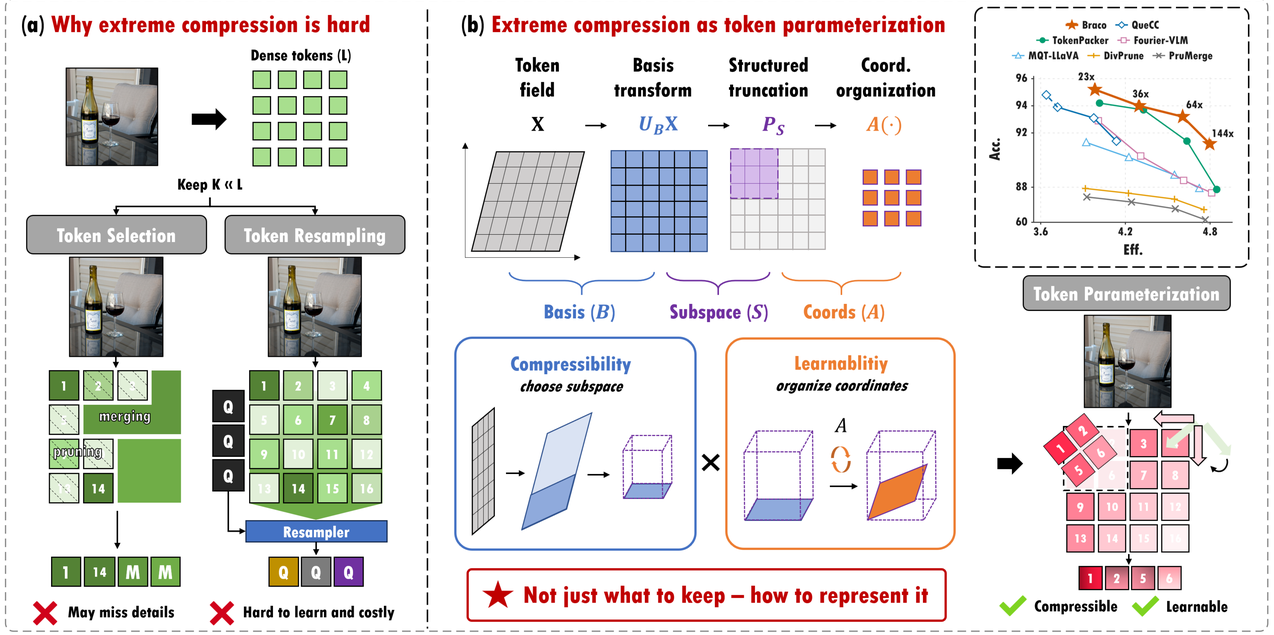
Contact: rzhong@zju.edu.cn · zyan2@zju.edu.cn

ZHEJIANG UNIVERSITY

NEURAL INFORMATION PROCESSING SYSTEMS

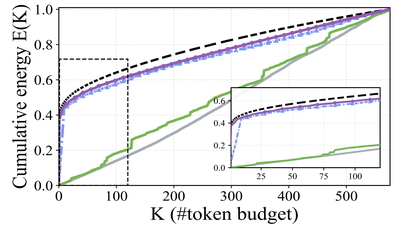
Token parameterization

*Braco formulates extreme visual-token compression as a **token parameterization problem**, jointly addressing **compressibility** and **learnability**.*



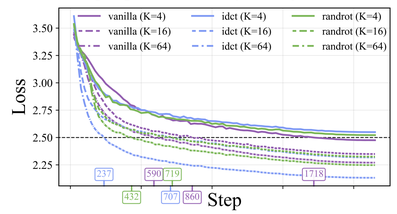
Compressibility

Preserve useful information:
energy retention + task readability.



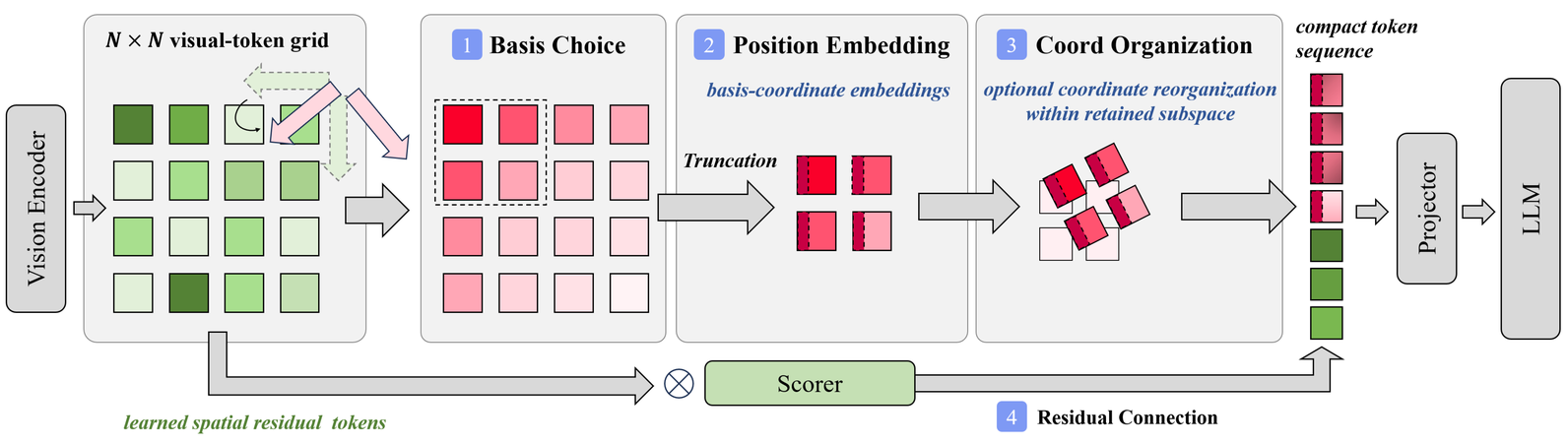
Learnability

Same subspace; different coordinates.
Conditioning + geometric compatibility.



Braco: a four-step coder

A compact transform-domain backbone + learned spatial residuals.



Main results

LLaVA-1.5-7B · 8 benchmarks

95.2%
Normalized accuracy
25 tokens · 23× compression

Method	GQA	MMB^{EN}	MMB^{CN}	MME^{All}	POPE^{F1}	SQA	VQA-T	MMVet	Acc. (%)	FLOPs (T)	Lat. (ms)
Upper Bound, 576 Tokens (1×)											
Vanilla	62.9	65.5	60.7	1785	85.7	69.7	58.0	32.8	100.0	8.67	67.25
25 Retained Tokens (23×)											
PruMerge (ICCV25)	49.8	55.8	47.0	1515	58.5	68.7	50.4	20.2	80.2	1.37	44.10
DivPrune (CVPR25)	54.0	59.8	50.8	1510	75.2	68.5	49.5	26.0	87.0	1.40	41.00
MQT-LLaVA (NIPS24)	57.1	61.4	53.1	1689	79.9	69.4	50.2	27.7	91.3	1.37	44.20
TokenPacker (IJCV25)	57.4	64.1	55.8	1688	**83.0**	69.4	**53.3**	29.3	94.2	1.38	38.40
Fourier-VLM	**58.4**	63.3	54.6	**1730**	**83.0**	69.7	50.2	27.5	92.9	1.37	39.83
Braco (ours)	58.1	**65.4**	**57.2**	1705	82.6	**70.0**	52.3	**30.4**	**95.2**	1.37	41.03
16 Retained Tokens (36×)											
PruMerge (ICCV25)	46.9	53.1	44.2	1446	51.6	68.2	49.3	18.8	76.2	1.25	43.37
DivPrune (CVPR25)	52.6	56.7	47.4	1411	71.3	67.3	48.3	24.8	83.2	1.28	40.26
MQT-LLaVA (NIPS24)	55.3	62.8	53.9	1630	78.2	68.9	48.2	27.6	90.2	1.25	44.27
QueCC (ICLR25)	**59.0**	63.1	54.6	1668	**83.5**	**70.6**	52.8	28.8	93.9	1.36	63.22
TokenPacker (IJCV25)	57.0	**63.7**	55.2	1681	83.0	69.3	**53.4**	29.0	93.7	1.26	37.87
Fourier-VLM	56.7	61.4	51.0	1644	82.5	69.6	48.5	27.2	90.3	1.25	40.00
Braco (ours)	57.5	63.3	**55.8**	**1704**	82.4	69.5	51.9	**29.8**	**94.0**	1.25	40.59
9 Retained Tokens (64×)											
PruMerge (ICCV25)	44.5	48.8	40.6	1352	47.0	66.5	46.1	15.8	70.7	1.15	42.90
DivPrune (CVPR25)	50.1	53.5	44.2	1320	67.8	65.8	46.2	22.0	78.5	1.17	39.91
MQT-LLaVA (NIPS24)	53.9	62.4	52.9	1565	78.8	69.5	47.1	27.1	88.9	1.15	43.02
QueCC (ICLR25)	**58.3**	62.9	**55.6**	1707	**83.3**	69.0	**51.4**	27.6	93.1	1.26	62.79
TokenPacker (IJCV25)	55.9	62.5	53.5	1645	81.8	68.7	51.2	27.7	91.4	1.16	37.40
Fourier-VLM	56.4	60.5	50.4	1683	81.4	67.4	47.0	24.9	88.5	1.15	39.88
Braco (ours)	57.0	**63.4**	54.2	**1773**	81.7	**69.7**	51.1	**28.2**	**93.2**	1.15	40.22
4 Retained Tokens (144×)											
PruMerge (ICCV25)	39.0	43.1	35.4	1210	38.5	62.9	40.2	12.3	62.0	1.09	42.60
DivPrune (CVPR25)	44.2	47.9	39.6	1199	60.0	62.6	40.8	18.2	70.2	1.11	39.50
MQT-LLaVA (NIPS24)	51.5	**62.9**	53.4	1449	78.2	69.8	46.2	25.8	87.1	1.09	44.82
QueCC (ICLR25)	**56.5**	61.0	53.5	1641	**81.7**	68.2	**50.4**	**28.8**	**91.4**	1.20	64.17
TokenPacker (IJCV25)	53.0	60.2	51.0	1550	79.0	66.8	47.0	24.0	86.2	1.10	37.10
Fourier-VLM	54.0	56.1	45.4	1609	81.0	64.8	45.7	21.5	83.5	1.09	40.30
Braco (ours)	54.4	62.2	**53.8**	**1658**	80.7	**69.9**	50.2	28.0	91.2	1.09	40.97

Compressor cost

vs QueCC · 16 tokens
Pre-projector boundary

Method	Boundary	Acc. (%)	Lat. (ms)	FLOPs (G)	Mem. (MB)
QueCC (ICLR25)	pre-proj.	93.9	17.864	109.504	14.875
MQT-LLaVA (NIPS24)	pre-proj.	90.2	5.154	2.674	11.336
PruMerge (ICCV25)	pre-proj.	76.2	3.300	0.724	13.749
Braco (ours)	pre-proj.	94.0	1.073	1.389	8.731
TokenPacker (IJCV25)	post-proj.	93.7	1.004	15.271	15.125
DivPrune (CVPR25)	post-proj.	83.2	1.061	29.603	23.008
Braco (ours)	post-proj.	94.0	1.270	2.060	8.731

16.6×
lower module latency

78.8×
fewer compressor FLOPs

Generalization

Normalized accuracy

Larger inputs · Vicuna-7B
98.1% 2,880 → 80 tokens, 36× compression

Smaller backbone · Qwen2.5-3B
90.8% 3,645 → 20 tokens, 182× compression

Explore & connect

arXiv · GitHub · Homepage · Xiaohongshu

Let's be friends