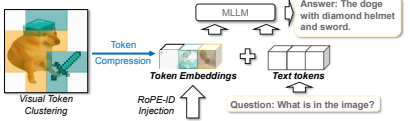
PPE: Positional Preservation Embedding for Token Compression in Multimodal Large Language Models

Mouxiao Huang*, **Borui Jiang***, Dehua Zheng, Hailin Hu, Kai Han, Xinghao Chen
Huawei Technologies

Paper | Code

Motivation

Token compression reduces redundancy but damages positional structure, → **breaking spatial layout in images and temporal continuity in videos**.

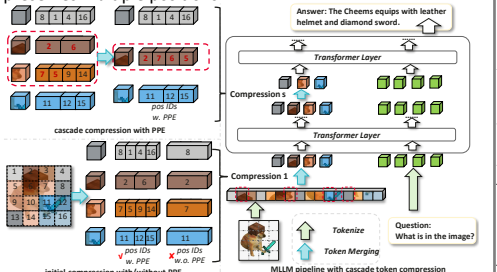


Key Contributions

- **Position Matters**: Preserves spatial-temporal structure under the same compression ratio.
- **Plug-and-Play**: No architecture change. No extra parameters.
- **Training-Free → Better with SFT**: Works out-of-the-box, improves further with training.
- **Cascade-Friendly**: Supports multi-stage compression for higher efficiency with minimal loss.
- **Across-the-Board Gains**: Consistent improvements, especially on layout-sensitive benchmarks.

PPE + Cascade Compression

PPE redistributes RoPE dimensions so each compressed token preserves multiple positions.

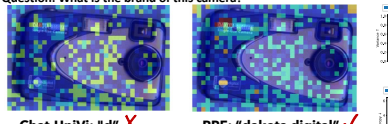


Key idea: different RoPE dimensions bind to different retained IDs

Plug-and-play | **No extra computation** | **Cascade within the LLM**

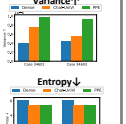
Visualization

Question: What is the brand of this camera?



Chat-UniVi: "d" ✗
Narrow attention misses the full brand string

PPE: "dakota digital" ✓
Broader positional coverage recovers the correct text



Main Results

55% token reduction | **+11.71** image avg gain | **+1.95** video avg gain

90% cascade reduction | **77%** valid IDs retained | **K=8** best trade-off

PPE vs. Full / Compressed Tokens

Model	VideoMME (w/o subs)	VideoMME (w subs)	MVBench	MMBench (EN)	MMBench (CN)	TextVQA	Reduction Ratio
LLVA-OneVision-0.5B	44.00	43.50	45.50	52.10	-	-	0%
InternVL2.5-4B	62.30	63.60	71.60	81.10	79.30	76.80	0%
Qwen2.5-VL-3B	61.50	67.60	67.00	79.10	78.10	79.30	0%
PACT-7B	57.60	-	-	80.30	-	75.00	67%
SparseVLM-7B	-	-	-	64.10	-	57.80	66%
PPE-3B (Ours)	**58.70**	**59.07**	**67.38**	**84.78**	**84.85**	**77.08**	55%
PPE*-3B (Ours)	58.48	58.52	67.35	-	-	-	90%

Training-Free & SFT

Setting	Benchmarks	Dense	Chat-UniVi	Chat-UniVi+PPE	VisionZip	VisionZip+PPE
Training-Free	MMBench (EN)	79.10	81.50	**82.28** (+0.78)	83.48	83.18 (-0.30)
	MMBench (CN)	78.10	80.06	**81.43** (+1.37)	81.75	81.64 (-0.11)
	TextVQA	79.30	37.60	**73.78** (+36.18)	79.00	79.62 (+0.62)
	DocVQA	93.90	19.58	**66.16** (+46.58)	83.98	**85.63** (+1.65)
	OCRBench	797	307	**598** (+291)	713	**725** (+12)
	ChartQA	84.00	18.72	**67.08** (+48.52)	79.72	**80.72** (+1.00)
SFT	MMBench (EN)	85.89	**84.92**	84.73 (-0.19)	**83.99**	83.78 (-0.21)
	MMBench (CN)	86.07	83.71	**84.87** (+1.16)	82.12	**82.63** (+0.15)
	TextVQA	79.50	57.66	**77.14** (+19.48)	80.02	82.06 (+2.04)
	DocVQA	89.44	52.48	**76.79** (+24.31)	84.84	90.52 (+5.68)
	OCRBench	761	535	**691** (+156)	711	**780** (+69)
	ChartQA	79.96	49.60	**74.52** (+24.92)	80.20	82.36 (+2.16)
	Reduction Ratio	0%	55%	55%	55%	55%

Performance vs. Efficiency

Method	MMBench (EN)	MMBench (CN)	TextVQA	Gen Time (s)	Memory (GB)	Reduction Ratio
PACT	74.14	74.17	73.73	0.08	15.82	89%
PACT + PPE	**74.48**	**75.00**	**73.87**	0.09	15.82	89%
ToMe	74.31	73.63	74.94	0.90	15.81	57%
ToMe + PPE	**74.57**	**74.74**	**76.16**	0.91	15.81	57%

K Controls Effective Positional Information

K	VideoMME (w/o subs)	VideoMME (w subs)	NeXT-QA (MC)	NeXT-QA (OE)	SEED-Bench (Video)	MVBench	Avg	Reduction Ratio	IDs Retained
1	57.74	58.04	77.88	28.77	55.07	**68.08**	57.97	55%	45%
8	**58.70**	**59.07**	**78.42**	**32.61**	55.98	67.38	**58.69**	55%	77%
24	58.19	58.56	78.02	31.64	55.52	67.73	58.48	55%	84%