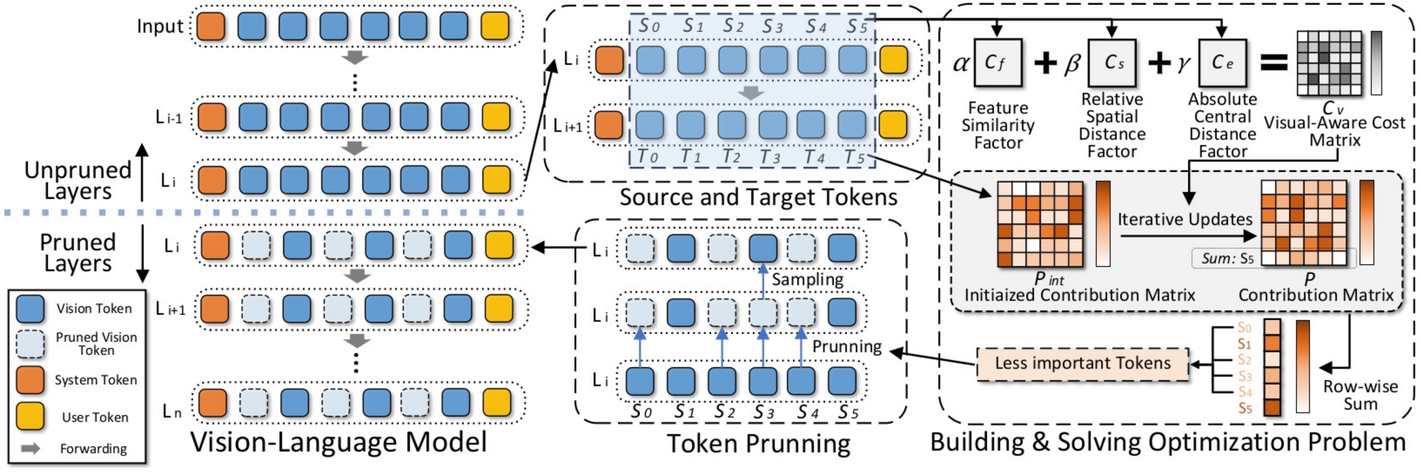
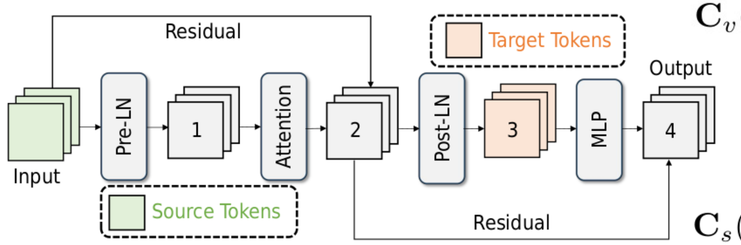
TopV: Compatible Token Pruning with Inference Time Optimization for Fast and Low-Memory Multimodal Vision Language Model

Cheng Yang , Yang Sui* , Jinqi Xiao , Lingyi Huang , Yu Gong , Chendi Li, Jinghua Yan,*
Yu Bai, Ponnuswamy Sadayappan, Xia Hu, Bo Yuan
Rutgers University, Rice University, The University of Utah, California State University, Fullerton

Overall Pipeline



Source & Target Tokens



Source & Target Tokens

$$\mathbf{C}_v(s_i,t_j) = \alpha\mathbf{C}_f(s_i,t_j)+\beta\mathbf{C}_s(s_i,t_j)+\gamma\mathbf{C}_e(s_i,t_j)$$

1. Feature Similarity Factor

$$\mathbf{C}_f(s_i,t_j) = \|s_i - t_j\|_F^2$$

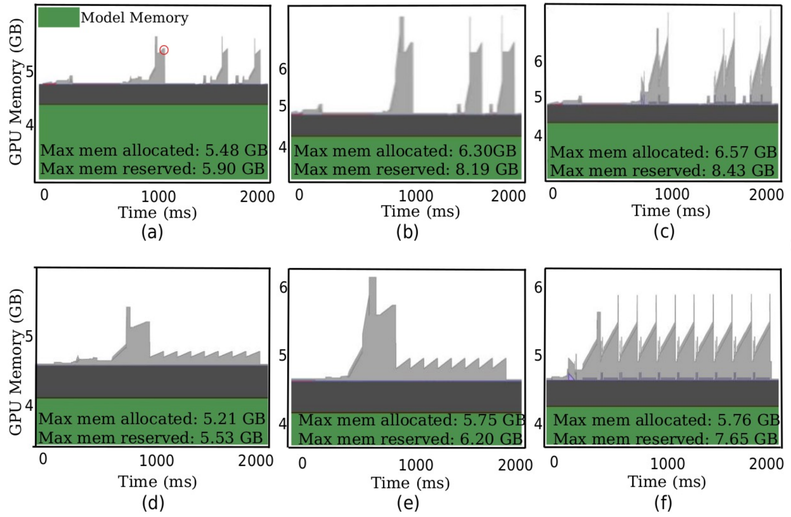
2. Relativea Spaatial Distance Factor

$$\mathbf{C}_s(s_i,t_j) = 1 - \exp\left(-\frac{(x_{s_i}-x_{t_j})^2+(y_{s_i}-y_{t_j})^2}{2\sigma^2}\right)$$

3. Absolute Central Distance Factor

$$\mathbf{C}_e(s_i,t_:) = \sqrt{(x_{s_i}-x_c)^2+(y_{s_i}-y_c)^2}$$

Memory Usage



Memory Usage Comparison
(InternVL-2B)

Tasks
AI2D (top)
OCRBench (bottom)

Methods
TopV (a, d)
Baseline (b, e),
FastV (c, f)

Next Token

Result

Model	Method	FLOPs Ratio ↑	AI2D Score↑	AI2D Mem.↓	AI2D Lat.↓	AI2D Tput.↑	SQA_IMG Score↑	SQA_IMG Mem.↓	SQA_IMG Lat.↓	SQA_IMG Tput.↑	MMMU Score↑	MMMU Mem.↓	MMMU Lat.↓	MMMU Tput.↑	MMBench Score↑	MMBench Mem.↓	MMBench Lat.↓	MMBench Tput.↑
LLaVA-v1.5-7B	Baseline	0	55.18	14.17	10'03	5.13	69.51	14.37	6'33	5.18	35.1	35.90	44'25	5.09	59.97	15.05	77'50	5.29
LLaVA-v1.5-7B	TopV	35%	55.41	13.98	9'25	6.07	69.56	14.3	6'02	6.18	35.4	35.9	41'52	5.76	60.42	14.9	72'36	6.4
LLaVA-v1.5-7B	FastV	47%	55.27	14.69	11'15	4.69	68.91	16.54	7'36	4.47	35.8	45.73	59'58	3.78	59.62	21.14	98'01	4.20
LLaVA-v1.5-7B	TopV	51%	55.31	13.87	8'30	6.12	69.61	14.17	5'32	6.27	35.1	35.45	39'30	5.82	59.65	14.77	67'02	6.46
LLaVA-v1.5-13B	Baseline	0	59.29	28.61	9'43	5.29	72.93	28.92	6'15	5.3	35	49.13	42'51	5.26	65.46	30.06	74'38	5.43
LLaVA-v1.5-13B	TopV	35%	59.38	28.27	8'15	6.25	73.27	28.66	5'22	6.26	35.2	47.79	36'51	6.11	65.63	29.41	63'58	6.32
LLaVA-v1.5-13B	FastV	48%	58.91	29.39	10'09	5.07	73.12	29.51	7'02	4.78	34.3	75.32	58'58	3.83	64.95	40.12	86'48	4.63
LLaVA-v1.5-13B	TopV	50%	59.27	27.91	7'58	6.45	73.17	28.32	5'07	6.57	34.7	47.29	35'21	6.38	65.16	29.11	60'23	6.56
InternVL2-2B	Baseline	0	72.67	6.39	9'10	5.61	94.2	6.63	6'07	5.5	34.56	13.72	39'57	5.25	67.6	6.78	86'57	12.1
InternVL2-2B	FastV	47%	71.57	6.92	11'16	4.64	93.75	7.69	7'45	4.34	33.67	32.65	52'46	3.98	69.46	7.62	136'56	7.68
InternVL2-2B	TopV	48%	73.35	5.57	7'50	6.57	94.22	5.99	4'50	6.39	34.6	13.33	36'20	5.93	69.7	5.98	75'17	14.57
InternVL2-26B	Baseline	0	83.13	56.79	41'57	2.46	97.47	57.61	28'43	2.34	47.11	69.02	173'47	2.24	80.89	57.02	342'27	4.24
InternVL2-26B	FastV	46%	81.27	57.67	49'05	2.1	96.82	58.12	33'12	2.02	46.91	165.23	230'12	1.69	80.86	58.87	490'38	2.96
InternVL2-26B	TopV	47%	83.34	55.91	34'03	2.86	97.67	56.73	24'26	2.75	46.98	68.26	157'36	2.47	81.27	56.33	300'52	4.83