Multi-Granular Spatio-Temporal Token Merging for Training-Free Acceleration of Video LLMs

Appendix

Tabs. A to D show the absolute values for the main comparison results.

Token Budget	Method	Q. Agn.	VNBench			VideoMME			LongVideoBench			MLVU			EgoSchema			NExT-QA			***Avg.***		
			Acc ↑	TTFT ↓	N_V ↓	Acc ↑	TTFT ↓	N_V ↓	Acc ↑	TTFT ↓	N_V ↓	Acc ↑	TTFT ↓	N_V ↓	Acc ↑	TTFT ↓	N_V ↓	Acc ↑	TTFT ↓	N_V ↓	Acc ↑	TTFT ↓	N_V ↓
100%	*LLaVA-Video 7B*	✓	77.6	0.962	11149	63.1	2.039	22086	59.6	1.805	19624	70.9	2.343	25088	58.7	2.312	25069	82.9	0.659	8116	68.8	1.687	18522
50%	+ FastV		72.7	0.503	5575	61.0	1.034	11043	57.4	0.918	9812	68.3	1.164	12544	57.6	1.166	12535	82.4	0.353	4058	66.6	0.856	9261
	+ DyCoke		72.1	0.458	5386	61.5	0.912	10555	58.1	0.820	9393	69.5	1.046	11978	58.6	1.049	11969	82.1	0.330	3957	67.0	0.769	8873
	+ FrameFusion		76.2	0.471	5529	62.0	0.961	10739	59.2	0.853	9616	69.4	1.083	12138	57.7	1.088	12298	82.6	0.336	4032	67.9	0.799	9059
	+ ToMe	✓	74.1	0.518	5575	61.4	1.043	11043	58.0	0.949	9812	69.7	1.192	12544	58.7	1.199	12535	82.6	0.370	4058	67.4	0.878	9261
	+ DyCoke-stage1	✓	73.2	0.495	5386	62.0	0.981	10555	58.2	0.877	9393	69.7	1.123	11978	58.7	1.118	11969	82.4	0.350	3957	67.4	0.824	8873
	+ STTM (Ours)	✓	77.4	0.455	4804	62.6	1.021	10771	59.6	0.895	9183	69.9	1.152	12187	58.6	1.045	10737	82.5	0.322	3452	68.4	0.815	8522
30%	+ FastV		61.6	0.336	3345	59.2	0.683	6626	54.7	0.610	5887	69.3	1.620	17562	56.9	0.771	7520	81.6	0.240	2435	63.9	0.710	7229
	+ DyCoke		55.9	0.308	3548	60.7	0.598	6878	57.1	0.544	6131	67.3	0.684	7798	58.3	0.693	7792	81.5	0.229	2631	63.5	0.509	5797
	+ FrameFusion		72.4	0.292	3157	60.7	0.581	6018	57.1	0.520	5462	67.5	0.650	6768	57.4	0.657	6870	81.9	0.218	2313	66.2	0.486	5098
	+ ToMe	✓	64.1	0.364	3345	59.2	0.720	6626	56.3	0.658	5888	67.0	0.821	7527	57.4	0.834	7521	81.6	0.265	2436	64.3	0.610	5557
	+ DyCoke-stage1	✓	63.9	0.358	3548	60.0	0.700	6878	56.5	0.631	6131	68.6	0.796	7798	58.7	0.800	7792	81.7	0.256	2631	64.9	0.590	5797
	+ STTM (Ours)	✓	76.0	0.299	2649	62.3	0.640	5929	57.0	0.616	5702	68.5	0.769	7337	58.0	0.773	7285	82.0	0.235	2168	67.3	0.555	5179

Table A. Comparison of training-free token reduction methods using LLaVA-Video-7B under 50% and 30% pre-filling token budgets.

Token Budget	Method	Q. Agn.	VNBench			VideoMME			LongVideoBench			MLVU			EgoSchema			NExT-QA			***Avg.***		
			Acc ↑	TTFT ↓	N_V ↓	Acc ↑	TTFT ↓	N_V ↓	Acc ↑	TTFT ↓	N_V ↓	Acc ↑	TTFT ↓	N_V ↓	Acc ↑	TTFT ↓	N_V ↓	Acc ↑	TTFT ↓	N_V ↓	Acc ↑	TTFT ↓	N_V ↓
100%	*LLaVA-OV 7B*	✓	68.8	0.922	11149	59.0	1.904	22086	56.3	1.712	19624	67.5	2.224	25088	61.3	2.216	25069	80.6	0.630	8116	65.6	1.601	18522
50%	+ FastV		65.6	0.467	5575	58.4	0.934	11043	56.2	0.842	9812	67.4	1.071	12544	61.0	1.080	12535	80.2	0.330	4058	64.8	0.787	9261
	+ DyCoke		64.9	0.418	5386	60.1	0.831	10555	57.9	0.740	9393	68.7	0.941	11978	61.8	0.948	11969	80.4	0.301	3957	65.6	0.697	8873
	+ FrameFusion		67.8	0.436	5568	59.8	0.865	10888	56.6	0.778	9719	68.2	0.994	12379	60.8	0.997	12511	80.8	0.312	4056	65.7	0.730	9187
	+ ToMe	✓	67.1	0.483	5575	59.8	0.963	11043	57.7	0.865	9812	68.8	1.101	12544	62.3	1.108	12535	80.1	0.343	4058	66.0	0.811	9261
	+ DyCoke-stage1	✓	65.1	0.459	5386	59.3	0.911	10555	58.2	0.821	9393	69.1	1.039	11978	61.8	1.049	11969	80.6	0.327	3957	65.7	0.767	8873
	+ STTM (Ours)	✓	71.2	0.404	4573	60.7	0.773	8579	57.7	0.804	9011	69.7	1.020	11692	61.7	0.972	10944	80.5	0.312	3628	66.9	0.714	8071
30%	+ FastV		59.2	0.303	3345	58.1	0.597	6626	55.3	0.543	5887	65.6	0.681	7526	60.9	0.694	7520	79.6	0.218	2435	63.1	0.506	5556
	+ DyCoke		53.9	0.280	3548	59.8	0.537	6878	55.9	0.488	6131	68.3	0.598	7798	62.1	0.607	7792	79.7	0.208	2631	63.3	0.453	5797
	+ FrameFusion		66.1	0.261	3263	59.0	0.499	6154	56.5	0.455	5566	66.7	0.570	6914	60.3	0.582	7251	79.9	0.195	2400	64.7	0.427	5258
	+ ToMe	✓	59.1	0.332	3345	59.8	0.649	6626	56.5	0.587	5888	69.1	0.740	7527	62.1	0.750	7521	79.5	0.240	2436	64.4	0.550	5557
	+ DyCoke-stage1	✓	58.4	0.324	3548	60.3	0.633	6878	56.2	0.577	6131	68.0	0.723	7798	62.3	0.731	7792	79.7	0.236	2631	64.1	0.537	5797
	+ STTM (Ours)	✓	70.4	0.277	2773	60.6	0.601	6264	56.6	0.570	6022	68.4	0.735	7989	61.8	0.570	5621	79.7	0.240	2577	66.2	0.499	5208

Table B. Comparison of training-free token reduction methods using LLaVA-OneVision-7B.

Token Budget	Method	VNBench			VideoMME			LongVideoBench			***Avg.***		
		Acc ↑	TTFT ↓	N_V ↓	Acc ↑	TTFT ↓	N_V ↓	Acc ↑	TTFT ↓	N_V ↓	Acc ↑	TTFT ↓	N_V ↓
100%	*Qwen2VL 7B*	66.4	2.438	22025	61.8	10.745	74982	56.8	10.597	72109	61.7	7.927	56372
50%	+ ToMe	63.4	1.130	11013	61.9	4.509	37491	56.7	4.348	36054	60.7	3.329	28186
	+ DyCoke-stage1	65.1	1.049	10645	62.6	4.166	36057	57.5	4.148	34735	61.7	3.121	27146
	+ STTM (Ours)	69.8	0.726	7228	62.9	4.761	39217	57.4	4.610	37315	63.4	3.366	27920
30%	+ ToMe	56.9	0.740	6608	61.4	2.835	22496	54.7	2.600	21633	57.6	2.058	16912
	+ DyCoke-stage1	54.2	0.706	6980	61.8	2.710	23479	57.5	2.694	22649	57.8	2.037	17703
	+ STTM (Ours)	66.7	0.420	6748	62.4	2.766	23143	56.9	2.472	20022	62.0	1.886	16638

Table C. Comparison using Qwen2VL-7B. Relative to 100% result.

Token Budget	Method	VideoMME		
		Acc ↑	TTFT ↓	N_V
100%	*LLaVA-Video 72B*	70.5	17.698	22086
50%	+ ToMe	70.6	8.424	11043
	+ DyCoke-stage1	70.4	8.052	10555
	+ STTM (Ours)	71.4	7.821	10082
30%	+ ToMe	68.5	5.186	6626
	+ DyCoke-stage1	69.3	5.353	6878
	+ STTM (Ours)	69.9	5.405	6897

Table D. Comparison using LLaVA-Video-72B. Relative to 100% result.