□

home

papers

study

life

About me

TS-LLaVA

- □□□□□TS-LLaVA: Constructing Visual Tokens through Thumbnail-and-Sampling for Training-Free Video Large Language Models
- □□□□□https://arxiv.org/abs/2411.11066
- arxiv

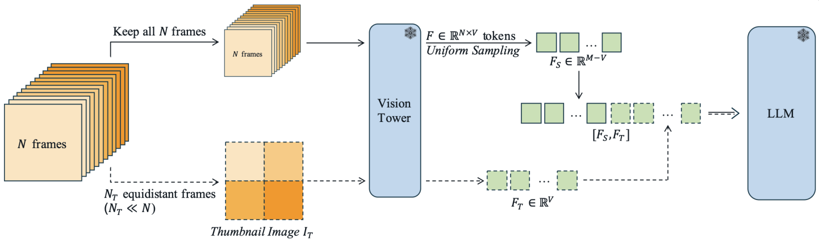

Figure 3. Illustration of our TS-LLaVA. The vision tower includes vision encoder and projection module in image LLM. The dashed lines and solid lines trace the procedures for constructing the thumbnail image tokens and sampled image tokens, respectively. V denotes the number of visual tokens from the vision tower, and M is the pre-defined number of visual tokens. We omit text input to LLM for simplicity.

□□□□□□□□□□□□□□□□□□□□□training-free□□□□□□□□□□□□□□□□□□□token□□SOTA□□□□□□□□□□□□□□□□

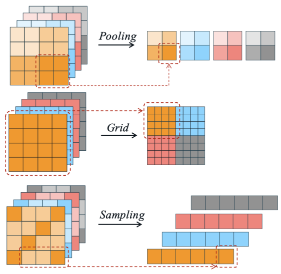

Figure 2. Visual token compression strategies illustrated. *Pooling* and *Sampling* operate on encoded tokens, *Grid* operates on RGB images. We omit the encoding procedure for simplicity. We extend *Grid* to *Grids* by composing multiple grid view images.

Compre. Strategy	#. of Frames (N)	#.of Vis. Tokens	Comp. Rate	Video-MME Short	Medium	Long	Overall
Concat	4	2304		49.2	41.0	36.9	42.4
Pooling	16	2304	4	49.6	42.6	36.4	42.9
Grid	4	576	4	47.0	39.8	36.2	41.0
Grids	16	2304	4	**52.2**	42.0	**37.6**	43.9
Sampling	16	2304	4	52.0	**43.0**	37.3	**44.1**

(a) Overall performance (Accuracy)

Compre. Strategy	#. of Frames (N)	#.of Vis. Tokens	Comp. Rate	Video-MME Reasoning Tasks Temporal	Spatial	Action	Object
Concat	4	2304		32.8	62.5	37.2	40.5
Pooling	16	2304	4	31.1	62.5	37.2	39.6
Grid	4	576	4	31.1	60.7	36.1	38.5
Grids	16	2304	4	**34.5**	**67.9**	37.2	**40.7**
Sampling	16	2304	4	33.3	66.1	**38.9**	40.5

(b) Performance (Accuracy) on reasoning tasks

Table 1. Comparison between different compression strategies on Video-MME dataset: Results obtained using LLaVA-v1.6-7B. "Vis. Tokens": Visual Tokens; "Comp.": Compression.

□□□6□□□□□□□□□□□□□□□□□□□□□□□□□□□□□□visionTower□□embedding□□□□□□□□□□□□□□VisionTower□□token□□□□□□□□□□□□□□□□□□□□□embedding□□□□□□□□□□□□□embedding□

Method	LLM Size	Vision Encoder	NExT-QA	EgoSchema	IntentQA
Video-LLaVA [20]	7B	ViT-L	60.5	37.0	-
Video-LLaMA2 [5]	7B	CLIP-L	-	51.7	-
MovieChat+ [36]	7B	CLIP-G	54.8	56.4	-
Vista-LLaMA [25]	7B	CLIP-G	60.7	-	-
DeepStack-L [28]	7B	CLIP-L	61.0	38.4	-
M^3 [3]	7B	CLIP-L	63.1	36.8	58.8
IG-VLM [13]	7B	CLIP-L	63.1	35.8	60.3
SF-LLaVA [45]	7B	CLIP-L	64.2	47.2	60.1
TS-LLaVA	7B	CLIP-L	**66.5**	**50.2**	**61.7**

(a) All models use 7B or comparable LLMs.

Method	LLM Size	Vision Encoder	NExT-QA	EgoSchema	IntentQA
Video-LLAMA2 [5]	46.7B	CLIP-L	-	53.3	-
LLoVi [48]	GPT-3.5	Unknown	67.7	50.3	64.0
VideoAgent [38]	GPT-4	Unknown	71.3	60.2	-
VideoTree [39]	GPT-4	Unknown	73.5	66.2	66.9
IG-VLM [13]	34B	CLIP-L	70.9	53.6	65.3
SF-LLAVA [45]	34B	CLIP-L	72.0	55.8	66.5
TS-LLaVA	34B	CLIP-L	**73.6**	57.8	**67.9**

(b) All models use 34B or stronger LLMs.

Table 3. Overall performance (Accuracy) obtained on *Multiple Choice VideoQA*. We **highlight** the top-performing training-free methods and underline the best-performing video LLMs overall. Methods below the dashed line (- -) are training-free, while those

above it have been trained on extensive video data.

□□□□□□□□□training-free□□□□□□□□□□□□□

- □□□training-free
- □□□multi-choice QA; MVBench; MLVU
- □□□1 A100
- □□□https://github.com/tingyu215/TS-LLaVA

Newer Older

2025-02-18
MIP-Adapter

2025-01-13
Thinking in Space