StreamForest: Efficient Online Video Understandingwith Persistent Event Memory

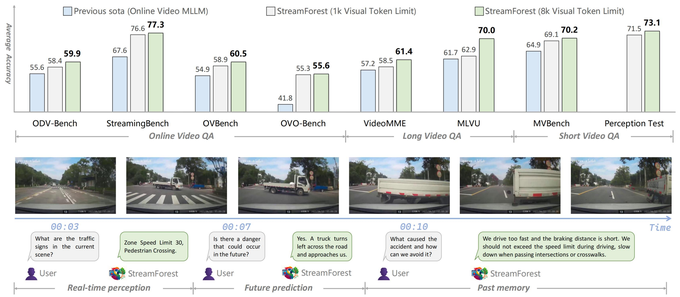
Xiangyu Zeng*[1,2], Kefan Qiu*[1], Qingyu Zhang*[1], Xinhao Li[1], Jing Wang[1], Jiaxin Li[1], Ziang Yan[3,2], Kun Tian[4], Meng Tian[5], Xinhai Zhao[4], Yi Wang[2], Limin Wang[1,2,†] (†: *corresponding author*)

[1]Nanjing University [2]Shanghai AI Laboratory [3]Zhejiang University
[4]Noah's Ark Lab, Huawei [5]Yinwang Intelligent Tech.

SCAN ME

NEURAL INFORMATION PROCESSING SYSTEMS

How to Help Models Remember What They See in Live Streams?

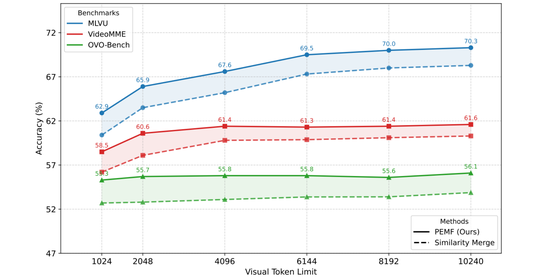


Understanding real-time live streams poses a much greater challenge than processing a complete pre-recorded movie. Traditional video models often forget what they just saw when handling continuously incoming streaming video.

In this study, we introduce an innovative dual-memory architecture that effectively overcomes this limitation. It enables the model to achieve both precise short-term perception and coherent long-term event memory.

Inspired by Human Memory: Smarter Streaming Video Understanding

Method	Size	Online Video			Long Video		Short Video	
		StreamingBench Real-Time All	OVBench Avg	OVO-Bench Overall	VideoMME w/o sub.	MLVU M-Avg	MVBench Avg	PerceptionTest Val
Open-source Offline Video MLLMs								
InternVL2 [9]	8B	63.7	48.7	50.1	54.0	64.0	65.8	-
LongVA [75]	7B	60.0	43.6	-	52.6	56.3	-	-
LLaVA-OneVision [28]	7B	71.1	49.5	52.9	58.2	64.7	56.7	57.1
Qwen2-VL [55]	7B	69.0	49.7	52.7	63.3	-	67.0	66.9
LongVU [50]	7B	-	-	48.5	60.6	65.4	66.9	-
LLaVA-Video [76]	7B	-	-	53.1	63.3	70.8	58.6	67.9
Open-source Online Video MLLMs								
VideoLLM-online [5]	8B	36.0	9.6	12.8	-	-	-	-
MovieChat [52]	7B	-	30.9	-	38.2	-	55.1	-
Flash-VStream [72]	7B	23.2	31.2	33.2	-	-	-	-
VideoChat-Online [23]	4B	-	54.9	-	52.8	-	64.9	-
Dispider [47]	7B	67.6	-	41.8	57.2	61.7	-	-
StreamForest	**7B**	**77.3**	**60.5**	**55.6**	**61.4**	**70.0**	**70.2**	**73.1**
StreamForest (FT-drive)	**7B**	**76.8**	**61.6**	**55.6**	**61.9**	**69.6**	**68.6**	**71.6**

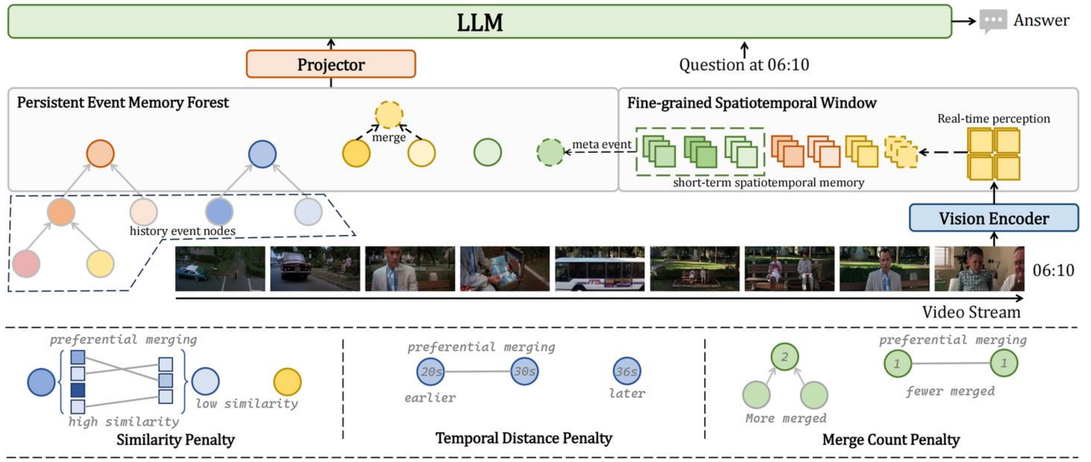


Remarkable results are observed in this study:

- SOTA Performance: The model achieves SOTA results across multiple mainstream online video understanding benchmarks, including StreamingBench, OVBench, and OVO-Bench. Moreover, under the streaming setting, it performs on par with the best offline video understanding methods on several offline benchmarks.

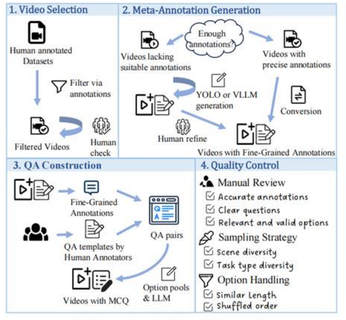
- Extreme Efficiency: Even under extreme visual compression (only 1024 tokens), the model retains 96.8% of the average accuracy compared to the full-budget setting, demonstrating remarkable efficiency and robustness.

StreamForest: Highly Efficient Model Architecture



ODV-Bench: A Comprehensive Online Video Understanding Benchmark

(a) Category Distribution of ODV-Bench



(b) Generation pipeline of ODV-Bench

(c) Task examples in ODV-Bench