StreamingBench: Assessing the Gap for MLLMs to Achieve Streaming Video Understanding

Junming Lin[1,3*] Zheng Fang[1*] Chi Chen[1] Haoxuan Cheng[3] Zihao Wan[1] Fuwen Luo[1]
Ziyue Wang[1] Peng Li[2] Yang Liu[1,2] Maosong Sun[1]

[1] Dept. of Comp. Sci. & Tech., Institute for AI, Tsinghua University
[2] Institute for AI Industry Research (AIR), Tsinghua University
[3] Beijing University of Posts and Telecommunications
[4] Xi'an Jiaotong University

Abstract

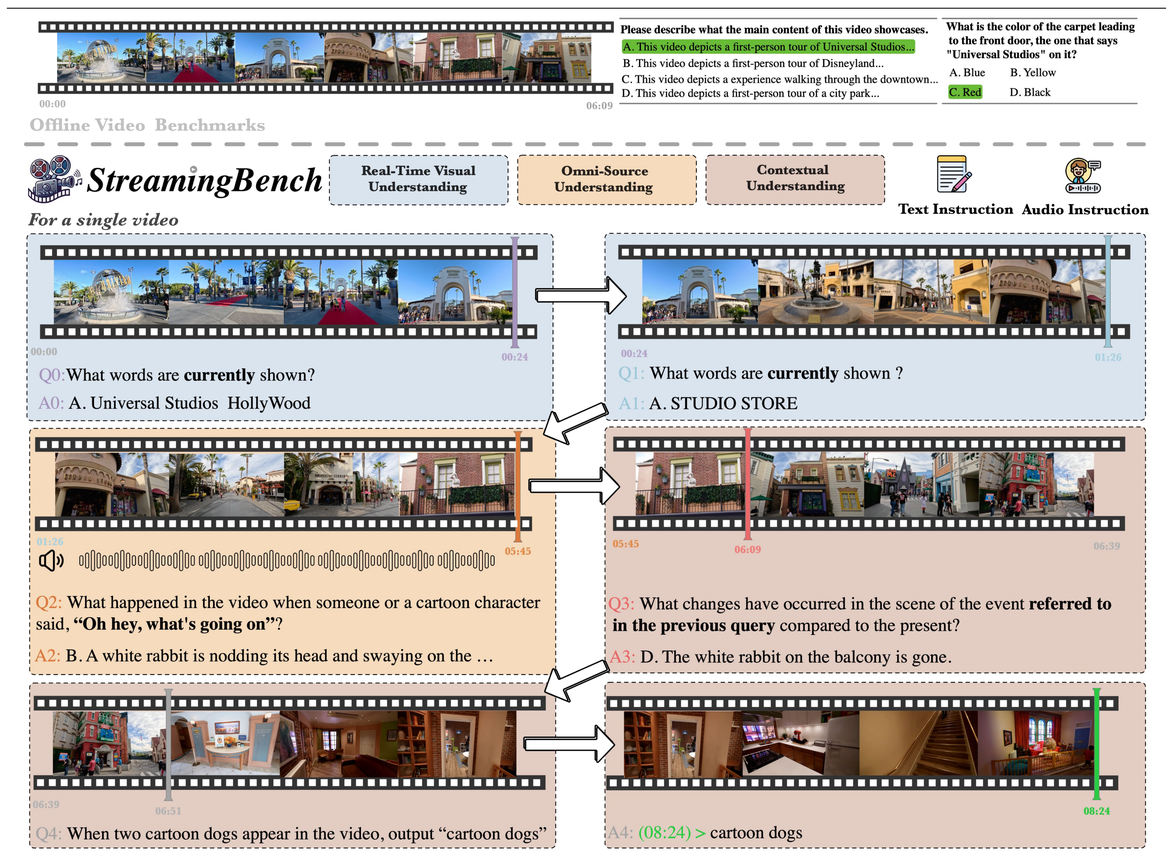
StreamingBench is the first comprehensive benchmark for **streaming video understanding** with timestamped, continuous, and audio-aware evaluation. It measures three capability groups: **real-time visual understanding**, **omni-source understanding**, and **contextual understanding**. The benchmark contains **900 videos**, **4,500 human-curated QA pairs**, and **18 tasks**. Experiments on **23 MLLMs** show that even the best proprietary model remains far below human performance, revealing major weaknesses in synchronized audio-visual reasoning, temporal grounding, contextual memory, and proactive response.

Motivation and Contributions

Streaming video understanding requires models to watch, listen, reason, and answer in real time. Existing video benchmarks are mostly offline: the full video is available before inference, timestamps are not central, and contextual interaction is weak.

- **StreamingBench** is designed for **timestamped streaming QA**, not offline replay.
- It evaluates **real-time visual**, **omni-source**, and **contextual** understanding in one benchmark.
- The benchmark contains **900 videos**, **4,500 human-curated QA pairs**, and **18 tasks** across diverse real-world categories.
- Results on **23 MLLMs** show a large gap between current systems and human-level streaming understanding.

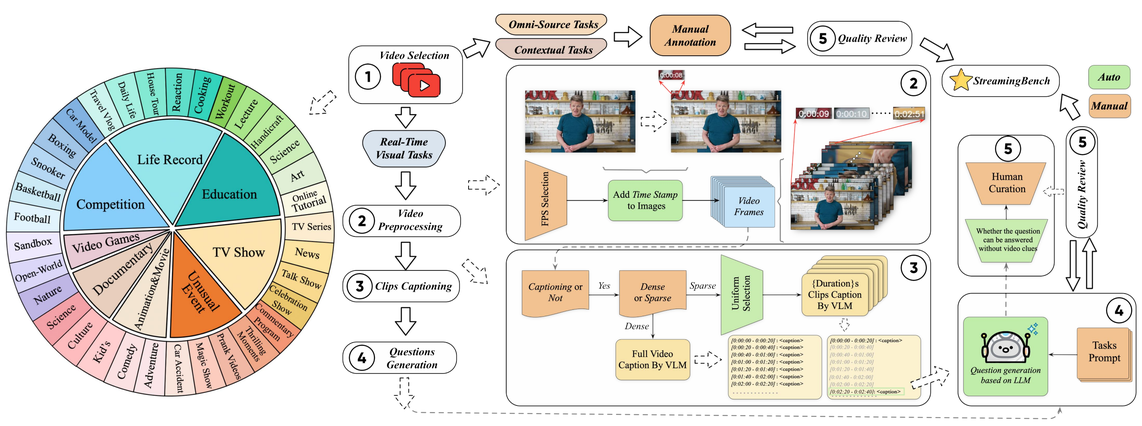
Streaming Setting vs. Offline Benchmarks



Each video is queried at multiple timestamps. Questions may depend on clues before, during, or after the query moment, and may require audio, memory, or proactive response.

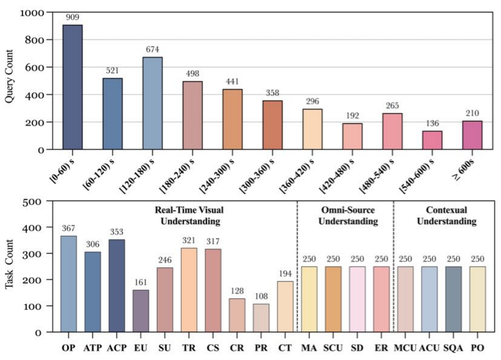
Benchmark Design and Data Construction

- **Real-Time Visual Understanding**: OP, CR, CS, ATP, EU, TR, PR, SU, ACP, CT.
- **Omni-Source Understanding**: ER, SCU, SD, MA with synchronized audio-visual evidence.
- **Contextual Understanding**: MCU, ACU, SQA, PO for misleading context, anomaly detection, multi-turn memory, and proactive output.



Videos are curated from eight categories. Real-time tasks are generated with timestamped captions and human verification; omni-source and contextual tasks are fully human-annotated.

Dataset Profile



Question counts span short and long videos, with balanced coverage across three capability groups.

Experimental Setup and Main Results

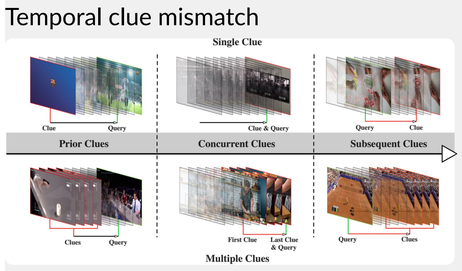
The paper evaluates **3 proprietary** and **20 open-source** MLLMs. For models without native streaming input, the evaluation uses the video context up to the query timestamp and supplies prior QA history for sequential tasks.

Model	Params	Frames	Real-Time Visual Understanding											Omni-Source Understanding					Contextual Understanding					Overall
			OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	All	ER	SCU	SD	MA	All	ACU	MCU	SQA	PO	All	
Human																								
Human‡	-	-	89.47	92.00	93.60	91.47	95.65	92.52	88.00	88.75	89.74	91.30	91.46	88.00	88.24	93.60	90.27	90.26	88.80	90.40	95.00	100	93.55	91.66
Proprietary MLLMs																								
Gemini 1.5 pro	-	video	**83.43**	**77.94**	**89.24**	81.65	**79.17**	**83.92**	83.93	60.32	**74.87**	**49.22**	**77.39**	52.40	**50.80**	**80.40**	**87.60**	**67.80**	**52.80**	42.40	**59.20**	45.10	**51.06**	**70.26**
GPT-4o	-	32	80.66	76.98	86.67	73.81	75.95	85.48	75.00	**70.66**	65.99	43.09	74.54	**53.60**	32.40	49.00	68.80	50.95	50.40	**42.80**	52.40	56.86	49.06	64.31
Claude-3.5-sonnet	-	20	82.45	73.77	82.43	**82.40**	76.39	85.56	61.68	60.73	67.88	47.62	74.04	39.60	35.60	34.40	56.00	41.40	36.00	43.20	34.80	**64.71**	39.70	60.06
Open-Source Video MLLMs ≈ 2B																								
LLaVA-OneVision	0.5B	32	**65.12**	61.72	**65.30**	**65.36**	56.52	**58.57**	57.41	**53.25**	**55.52**	33.16	**58.28**	37.20	24.40	29.20	34.00	31.20	29.20	**30.40**	28.40	9.80	28.09	**46.36**
Qwen2-VL	2B	1fps	61.85	**68.75**	60.57	63.73	56.52	58.26	**60.19**	50.00	50.42	37.82	56.76	36.40	23.20	27.20	36.40	30.80	24.40	23.20	**40.80**	5.88	27.96	45.36
InternVL2	2B	32	56.95	50.78	58.04	64.71	**57.14**	52.65	50.93	44.72	46.46	**47.67**	53.52	**32.00**	**30.00**	**29.20**	**43.60**	**33.70**	**32.40**	26.80	37.20	**13.73**	**30.96**	44.71
Open-Source Video MLLMs ≈ 7B																								
MiniCPM-o 2.6	7B	64	**85.01**	**85.94**	**89.91**	85.95	80.12	**86.60**	76.85	**74.80**	**74.79**	46.11	**79.88**	48.40	24.40	**63.20**	**77.60**	**53.40**	38.00	36.80	42.40	29.41	38.45	**66.01**
InternVL2.5	8B	32	82.02	75.00	85.80	**88.24**	**81.99**	81.93	75.00	73.98	74.50	**50.78**	78.32	**50.00**	**30.40**	50.40	56.00	46.70	**40.40**	**38.00**	**50.00**	**43.13**	**43.14**	64.36
InternLM-XComposer2.5-OmniLive	7B	64	84.74	79.69	83.91	87.91	75.78	80.06	67.59	65.85	70.25	38.34	75.36	45.20	**30.40**	48.40	60.80	46.20	38.00	28.80	38.00	13.73	33.58	60.80
Qwen2.5-VL	7B	1fps	77.66	75.00	84.54	83.66	76.40	79.13	**85.19**	65.85	68.27	46.11	74.64	43.60	24.80	44.00	57.60	42.50	35.20	25.60	48.40	23.53	30.70	59.89
LLaVA-OneVision	7B	32	82.83	77.34	83.23	83.33	72.05	74.77	73.15	68.29	71.10	41.97	74.27	41.20	26.00	43.20	52.80	40.80	34.80	30.40	30.00	29.55	31.59	58.54
MiniCPM-V 2.6	8B	32	78.20	71.88	84.18	83.99	75.16	75.39	72.22	56.50	67.14	47.15	72.43	42.00	27.60	40.40	50.80	40.20	37.20	27.20	40.00	22.22	34.00	57.78
VITA-1.5	7B	1fps	77.66	82.81	82.33	79.34	72.67	71.34	67.59	62.20	69.12	32.12	70.88	42.80	28.40	39.60	52.40	40.80	32.00	33.60	44.00	25.49	35.83	57.36
InternVL2	8B	32	73.84	65.63	78.80	82.03	71.43	72.90	73.15	63.01	65.44	42.49	70.11	44.80	28.00	47.20	50.80	42.70	34.80	27.20	42.80	40.91	35.31	57.26
Qwen2-VL	7B	1fps	75.75	79.69	76.58	79.08	74.53	75.08	74.07	65.85	65.16	41.97	71.15	40.80	25.20	41.20	55.60	40.73	34.00	26.40	44.40	22.73	34.16	57.19
LLaVA-NeXT-Video	32B	32	80.11	71.09	80.70	80.72	71.43	73.21	62.96	59.35	63.17	36.79	69.83	41.60	24.40	44.40	56.40	41.73	34.00	28.80	44.00	18.18	34.49	56.71
Kangaroo	7B	64	77.57	74.22	75.70	74.38	70.91	62.86	52.63	50.54	63.97	33.16	65.76	40.80	34.00	40.00	45.20	40.04	35.60	32.40	30.80	16.00	31.86	53.46
Ola	8B	64	61.26	71.09	68.25	60.20	67.72	57.94	74.07	53.66	58.69	20.73	58.80	41.60	28.00	27.20	42.00	34.70	32.00	26.00	33.20	18.18	27.35	51.84
LongVA	7B	128	73.02	66.41	66.46	74.84	65.22	62.93	60.19	56.91	56.66	37.82	63.11	41.20	28.40	32.00	42.00	35.90	33.20	28.00	34.40	15.91	30.85	50.78
VILA-1.5	8B	16	71.12	57.81	74.68	72.22	70.81	62.62	51.85	51.22	60.34	18.65	61.54	40.40	28.40	37.20	44.00	37.53	28.80	28.00	26.00	17.65	26.97	49.51
LongVILA	8B	32	67.85	49.22	60.25	71.57	56.52	56.70	40.74	56.50	57.51	41.97	58.48	37.20	25.20	30.40	40.40	33.30	29.20	28.80	25.60	7.84	26.59	46.60
Video-LLaMA2	7B	32	59.95	60.16	62.97	60.46	54.66	46.11	41.67	46.75	48.16	34.72	52.58	43.60	23.20	35.20	41.60	35.92	28.00	26.00	21.20	0	23.47	43.28
Video-CCAM	14B	32	54.50	63.28	73.73	63.73	57.76	47.35	50.93	41.87	48.44	26.94	53.42	38.00	21.20	31.20	38.40	32.22	26.00	22.80	27.60	22.73	25.29	43.25
Oryx	7B	32	51.50	48.44	57.10	52.50	59.63	41.74	37.04	48.78	48.16	13.99	47.20	31.60	23.20	19.20	40.80	28.70	29.60	25.20	33.20	5.88	27.84	39.29
VideoLLM-online	8B	64	36.71	43.75	39.23	33.99	46.43	36.45	39.81	34.78	39.71	24.35	37.06	34.80	21.60	26.80	34.40	29.40	23.20	33.20	31.60	15.69	28.46	33.68
LongLLaVA	7B	32	27.32	43.75	39.87	36.39	30.63	28.35	25.92	29.80	26.35	39.38	32.18	32.40	24.80	25.60	28.00	27.70	24.00	24.40	33.20	7.84	25.97	29.98
Flash-VStream	7B	64	28.88	28.13	25.65	26.47	31.68	23.36	25.00	24.39	27.48	25.91	26.58	32.80	23.60	26.00	30.40	28.20	25.20	26.40	29.20	3.92	25.47	26.75
Open-Source Video MLLMs ≈ 72B																								
LLaVA-OneVision	72B	32	**79.56**	**85.94**	**82.97**	**83.28**	**78.75**	**74.77**	**81.48**	**67.87**	**73.37**	**51.30**	**75.98**	**44.80**	26.80	**43.60**	**55.20**	**42.60**	**40.40**	**44.80**	36.40	**21.57**	**39.33**	**61.39**
Qwen2-VL	72B	1fps	75.48	82.03	79.18	79.41	74.53	71.03	76.85	64.63	69.69	49.22	72.28	39.20	**36.00**	33.60	53.60	40.60	31.60	38.40	**40.80**	9.80	35.21	58.01
VideoLLaMA2	72B	32	64.85	75.78	77.29	72.79	66.25	58.88	59.26	53.66	58.07	48.19	63.70	43.20	24.40	35.20	49.60	38.10	38.80	30.00	36.80	13.73	33.83	52.19

- Best proprietary model: **Gemini 1.5 Pro, 70.26%**, still **21.4 points below human**.
- Best open-source model: **MiniCPM-o 2.6, 66.01%**.
- Models are strongest on **real-time visual** tasks and weakest on **contextual** and **omni-source** tasks.
- **PO** and **SQA** remain especially difficult, indicating weak temporal precision and long-horizon interaction memory.

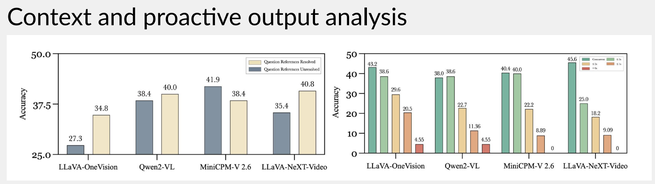
Why Models Still Struggle

Temporal clue mismatch



Questions are categorized by whether the minimal clue is **prior**, **concurrent**, or **subsequent**, and whether it is **single** or **multiple**. Performance drops sharply on concurrent and subsequent clues.

Context and proactive output analysis



- Making references explicit improves **SQA**, but only modestly.
- Converting **PO** into a concurrent query improves scores substantially.
- The main bottlenecks are **context tracking**, **time alignment**, and **following proactive instructions**.

Streaming Prefill and Audio Input Matter

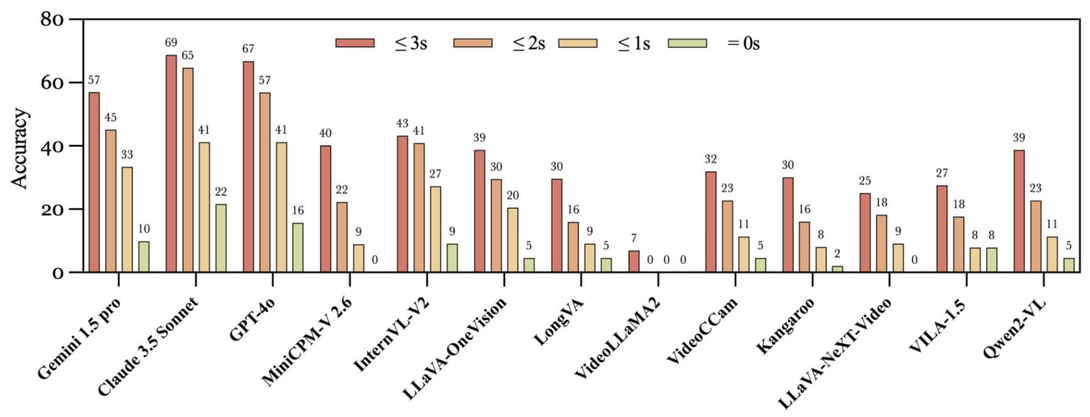
Model	Input Frames	Offline		Stream	
		TTFT (s)	Latency (s)	TTFT (s)	Latency (s)
Audio Instruction					
MiniCPM-o 2.6	64	6.32	35.45	0.71	9.60
IXC2d5-OL + ASR	32	8.43	75.50	-	-
Ola	64	2.19	14.10	-	-
VITA-1.5	64	1.10	26.50	-	-
Text Instruction					
MiniCPM-o 2.6	64	6.12	34.60	0.13	5.90
VideoLLM-Online	64	-	-	0.10	7.25
IXC2d5-OL	32	7.58	95.15	-	-
Ola	64	2.07	14.75	-	-
VITA-1.5	64	1.12	27.30	-	-
Qwen2.5-VL	64	4.84	28.05	-	-
InternVL-2.5	32	7.76	39.70	-	-

Latency and TTFT comparison of various MLLMs across different interaction modes.

Model	w/o ASR	w/ ASR	Δ
LLaVA-NeXT-Video	41.73	48.10	+6.37
LLaVA-OneVision	40.80	47.20	+6.40
Qwen2-VL	40.73	47.20	+6.47
MiniCPM-V 2.6	40.20	44.40	+4.20
InternVL2	42.70	42.60	-0.10
LongVA	35.90	39.50	+3.60
Kangaroo	40.04	42.50	+2.46
Video-CCAM	32.22	36.70	+4.48

Performance on Omni-Source Understanding with ASR and improvement over non-ASR variants.

Proactive Interaction Ability Matter



Performance of different MLLMs on the proactive output task. "≤ xs" means that the answer is considered correct if the actual output time is within x seconds of the ground truth.

Takeaways

- StreamingBench moves evaluation from offline video QA to **continuous, streaming, proactive interaction**.
- Current MLLMs are still far from human performance, especially when **audio**, **context**, and **precise temporal grounding** are required together.
- Future progress likely depends on **native streaming architectures**, stronger **audio-visual fusion**, and better **episodic memory / proactive interaction**.