

Table 1. We complement the ablation on memory propagation and temporal selection and add explanations.

Memory Selection EgoSchema Global Breakpoint Explanation					
✗	✗	34.4	52.5	21.6	Uniform sampling 64 frames
✓	✗	37.3	69.1	23.0	No propagated memory, concatenate time indicators with questions for selection
✗	✓	38.4	43.8	39.1	No temporal selection, adopt memories of final 4 clips
✓	✓	44.1	90.4	54.9	Full architecture of VideoStreaming

User: What challenges does the protagonist in the movie face to achieve his goal?

Text-only LLM: I don't see a specific movie mentioned in your question. Could you please provide the title of the movie you're referring to?



VideoStreaming: The protagonist in the movie confronts a series of obstacles in his endeavor to accomplish a remarkable feat. He must devise a method to access the towering structures, tackle the complexities of setting up the wire, maintain his equilibrium and concentration while suspended on a tightrope, and manage the risks associated with installing the equipment on the towers.

Figure 1. Visualization of the grounding results and comparison of text-only LLM and VideoStreaming with the reformulated question.

Table 2. We compare using ViT based model and the language model Phi with different layers as the streaming encoder.

Encoder Layers	Params	EgoSchema	Next-QA	Global Breakpoint		
MC-ViT	24	0.4B	32.3	53.1	71.2	40.4
Phi	4	0.3B	36.4	59.6	77.3	46.2
Phi	8	0.7B	39.8	63.2	84.3	49.2
Phi	12	1.0B	42.5	65.1	87.4	51.2
Phi	16	1.3B	44.1	66.2	90.4	54.9
Phi	24	2.0B	43.8	66.0	90.0	53.7

Table 3. We compare with LLaMA-VID on EgoSchema, Next-QA and MovieChat-1K benchmarks.

Method	Params	EgoSchema	Next-QA	Global Breakpoint	
LLaMA-VID	13B	35.5	50.3	62.3	46.1
Ours	7B+1.3B	44.1	66.2	90.4	54.9

Table 4. We compare with VideoChat2 on EgoSchema, MLVU and VideoMME medium and long subset (denoted as VideoMME-M and VideoMME-L).

Method	EgoSchema	MLVU	VideoMME-M	VideoMME-L
VideoChat2	54.4	44.5	37.0	33.2
Ours	48.1	51.7	43.1	39.6

Table 5. We compare using different layers of Phi and Vicuna as the streaming encoder.

Encoder Layers	Params	EgoSchema	Next-QA	Global Breakpoint		
Phi	8	0.7B	39.8	63.2	84.3	49.2
Vicuna	3	0.7B	36.6	58.4	77.8	46.3
Phi	16	1.3B	44.1	66.2	90.4	54.9
Vicuna	6	1.3B	39.5	64.1	85.5	50.1

Table 6. We ablate different settings, including memory propagation, temporal selection, the number of summarization tokens and the number of selected clips, on hour-long MovieNet-QA benchmark from three perspectives.

Memory	Selection	P	V	Overview	Plot	Temporal
✗	✗	-	-	1.98	2.23	1.39
✓	✗	16	4	2.51	2.61	1.63
✗	✓	16	4	2.24	2.77	1.52
✓	✓	16	4	2.65	3.13	1.88
✓	✓	4	4	2.53	2.82	1.73
✓	✓	4	8	2.58	3.02	1.83
✓	✓	16	8	2.68	3.17	1.95

Table 7. We compare the performance and average number of sampled frames with clip-based and frame-based sampling.

Sampling	Metric	EgoSchema	Global Breakpoint	MovieNet-Avg
Clip	Acc	44.1	90.4	54.9
	Frames	176	464	464
Frame	Acc	44.0	90.7	54.2
	Frames	180	460	460

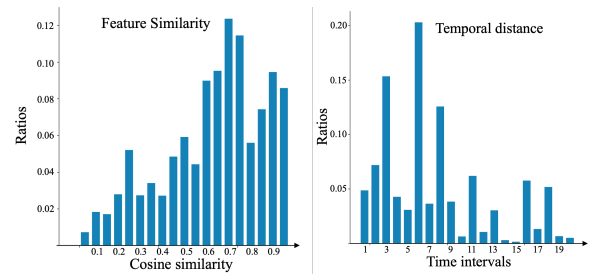


Figure 2. Visualization of the feature similarity and temporal distance of selected clips.