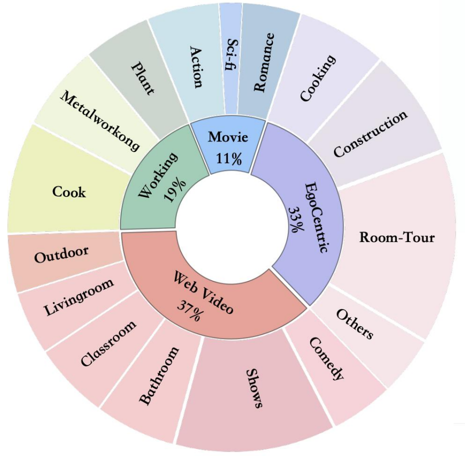
ICLR
International Conference On Learning Representations

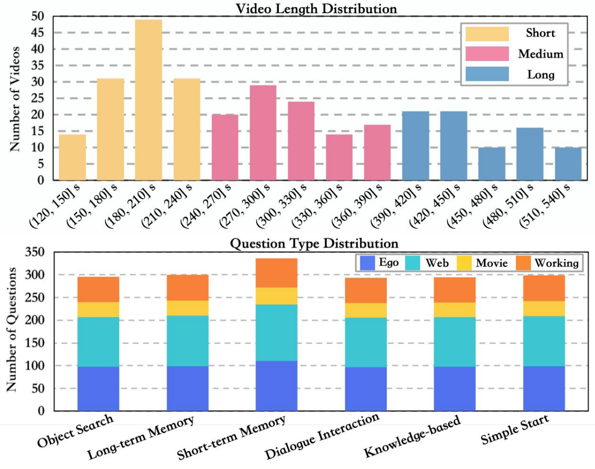
Streaming Video Understanding and Multi-Round Interactions with Memory-Enhanced Knowledge

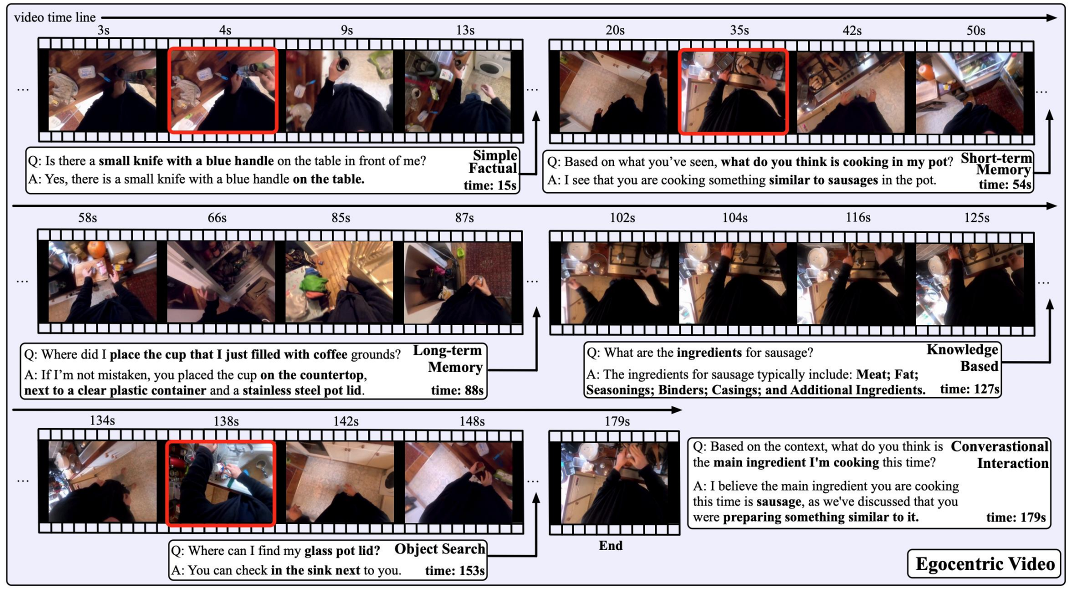
Haomiao Xiong , Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, Huchuan Lu

Introduction of the StreamBench

- **1.8K** manually annotated high-quality question-answer pairs.
- **306** video data with **16** subcategories, average duration **4.5** mine.
- **6 types of questions**: long-term memory, short-term memory, object search, interaction, knowledge based, and factual content.







Q: Is there a **small knife with a blue handle** on the table in front of me?
A: Yes, there is a small knife with a blue handle **on the table.**
Simple Factual **time: 15s**

Q: Based on what you've seen, **what do you think is cooking in my pot?**
A: I see that you are cooking something **similar to sausages** in the pot.
Short-term Memory **time: 54s**

Q: Where did I **place the cup that I just filled with coffee** grounds?
A: If I'm not mistaken, you placed the cup **on the countertop, next to a clear plastic container** and a **stainless steel pot lid**.
Long-term Memory **time: 88s**

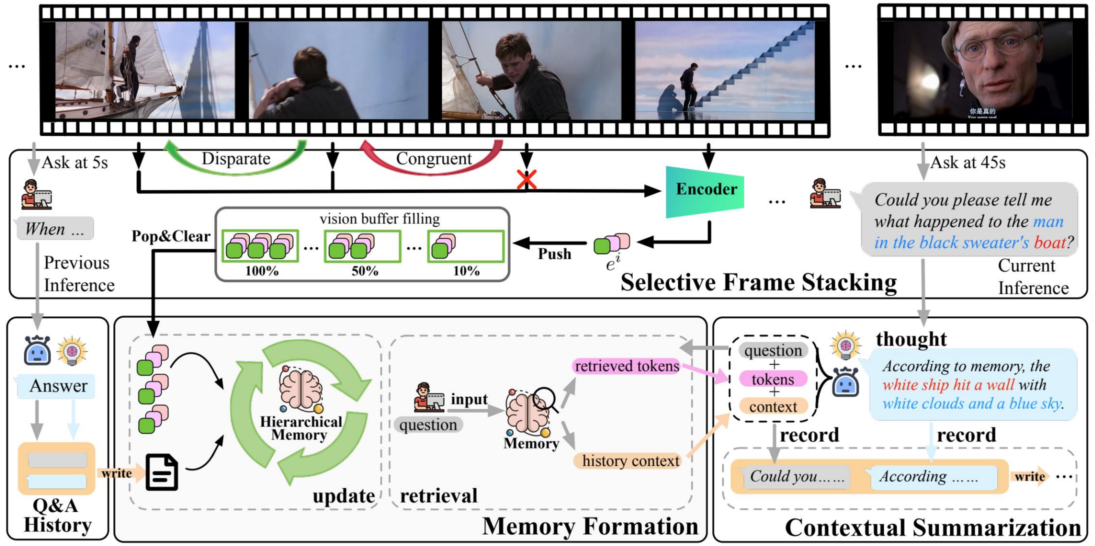
Q: What are the **ingredients** for sausage?
A: The ingredients for sausage typically include: **Meat; Fat; Seasonings; Binders; Casings; and Additional Ingredients.**
Knowledge Based **time: 127s**

Q: Where can I find my **glass pot lid?**
A: You can check **in the sink next** to you.
Object Search **time: 153s**

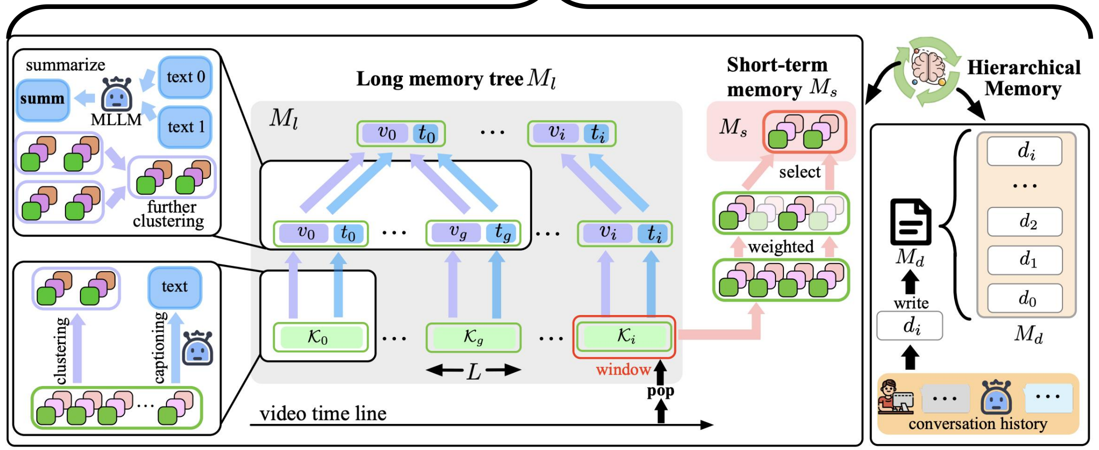
Q: Based on the context, what do you think is the **main ingredient I'm cooking** this time?
A: I believe the main ingredient you are cooking this time is **sausage**, as we've discussed that you were **preparing something similar to it.**
Converastional Interaction **time: 179s**

Egocentric Video

StreamChat with Hierarchical Memory Storage



- **Long-term Memory**: Storage compressed video features.
- **Short-term Memory**: Contextual information supplement.
- **Dialogue Memory**: Historical dialogue information.

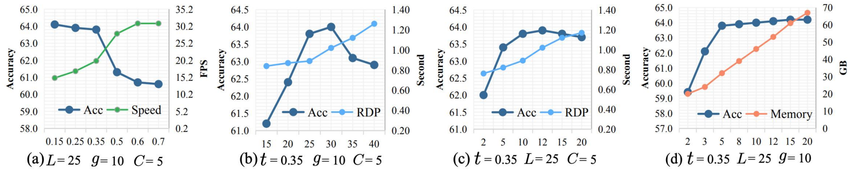


(a) Long-short term Memory (b) Dialogue Memory

Balanced Performance on Online and Offline Scenarios

1. StreamBench: 3.48 sco, and 64.7% acc.

Method	Publication	OS		LM		SM		CI		KG		SF	
		Sco.	*Acc.*	*Sco.*	*Acc.*	*Sco.*	*Acc.*	*Sco.*	*Acc.*	*Sco.*	*Acc.*	*Sco.*	*Acc.*
Human performance	- -	3.95	71.8	3.81	69.3	4.07	81.5	4.14	82.6	4.06	80.7	4.30	80.7
GPT-4o-50 [27]	Arxiv 2024	3.27	60.5	3.35	61.2	3.41	64.4	3.81	72.3	4.58	93.9	3.83	74.7
GPT-4o-35 [27]	Arxiv 2024	3.22	59.6	3.28	58.6	3.45	65.3	3.76	71.7	4.54	93.3	3.50	66.1
GPT-4o-mini-35 [27]	Arxiv 2024	2.52	46.8	2.70	45.8	2.80	51.0	3.50	64.0	4.67	95.2	2.90	53.3
Instruct-tuning													
Video-LLaVA [4]	EMNLP 2024	2.25	31.2	2.31	35.9	2.50	41.8	3.18	56.1	3.81	74.6	2.93	54.8
LLaMA-VID [2]	ECCV 2024	2.32	33.9	2.43	38.2	2.63	44.1	3.31	58.4	3.93	76.9	3.06	57.1
VILA1.5 [31]	CVPR 2024	2.33	36.1	2.54	44.3	2.87	50.8	3.59	68.3	3.97	78.6	3.38	65.5
InternVL2 [32]	CVPR 2024	2.49	38.5	2.70	46.6	2.89	50.9	3.61	67.6	4.02	81.0	3.29	62.2
LLaVA-NExT [28]	Arxiv 2024	2.17	35.0	2.14	31.4	2.15	36.0	2.55	42.7	3.88	76.1	3.12	57.6
LLaVA-Hound [29]	Arxiv 2024	2.49	37.6	2.68	43.2	3.09	53.4	3.21	55.7	3.89	76.3	3.35	62.0
LongVA [20]	Arxiv 2024	2.61	41.8	2.81	47.4	3.20	57.6	3.29	59.8	4.01	80.7	3.48	66.1
MiniCMP-v2.6 [30]	Arxiv 2024	2.32	37.6	2.78	51.9	2.62	43.7	3.35	65.7	3.19	66.2	3.27	64.2
InternLM-XCP2.5 [33]	Arxiv 2024	2.40	38.8	2.81	43.3	2.89	50.8	3.62	65.6	4.41	88.4	3.23	60.5
Training-Free													
MovieChat [7]	CVPR 2024	1.45	18.6	1.42	20.4	1.76	26.5	2.28	42.3	3.39	67.2	2.05	35.8
FreeVA [8]	Arxiv 2024	2.39	35.6	2.33	37.5	2.62	43.7	3.16	58.8	4.24	84.0	2.87	53.7
Online													
Video-online [11]	CVPR 2024	2.61	41.4	2.87	48.8	3.01	52.9	3.31	62.7	3.58	69.2	3.39	64.1
Flash-VStream [10]	Arxiv 2024	2.38	37.1	2.64	44.5	2.78	48.6	3.13	58.1	3.34	66.4	3.17	59.2
STREAMCHAT													
Slow	- -	**3.01**	**51.7**	**2.93**	**53.9**	**3.21**	**57.8**	**3.86**	**68.5**	**4.38**	**88.1**	**3.57**	**69.3**
Base	- -	2.93	50.5	2.87	52.9	3.15	56.1	3.82	67.6	4.37	87.9	3.56	68.8
Fast	- -	2.78	48.1	2.73	49.5	3.02	53.5	3.69	65.2	4.12	86.7	3.46	67.6



(a) $L=25$ $g=10$ $C=5$ (b) $t=0.35$ $g=10$ $C=5$ (c) $t=0.35$ $L=25$ $C=5$ (d) $t=0.35$ $L=25$ $g=10$

M_l	M_s	M_d	OS		LM		SM		CI		KG		SS		Average	
			Sco.	*Acc.*	*Sco.*	*Acc.*	*Sco.*	*Acc.*	*Sco.*	*Acc.*	*Sco.*	*Acc.*	*Sco.*	*Acc.*	*Sco.*	*Acc.*
✗	✗	✗	2.54	41.6	2.55	45.5	2.93	52.5	3.30	60.1	**4.44**	**89.9**	**3.79**	**72.6**	3.27	60.3
✗	✗	✓	2.55	41.9	2.55	45.7	2.94	52.5	3.66	64.2	**4.44**	88.7	3.78	72.4	3.32	60.9
✗	✓	✗	2.58	43.3	2.62	46.6	3.09	55.7	3.31	60.7	4.39	88.1	3.68	69.8	3.28	60.7
✓	✗	✗	2.85	49.5	2.78	51.7	2.96	53.5	3.32	61.1	4.42	88.4	3.65	69.4	3.33	62.2
✓	✓	✗	2.91	50.4	**2.88**	**53.0**	3.10	56.0	3.55	63.4	4.36	87.6	3.58	68.7	3.39	63.1
✓	✓	✓	**2.93**	**50.5**	2.87	**52.9**	**3.15**	**56.1**	**3.82**	**67.6**	4.37	87.9	3.56	68.8	**3.42**	**63.8**

2. Offline Benchmarks: 2.77 sco, and 50.6% acc.

Method	Publication	ActNet		NExT-QA		MSVD		MSRVTT		Average	
		Sco.	*Acc.*	*Sco.*	*Acc.*	*Sco.*	*Acc.*	*Sco.*	*Acc.*	*Sco.*	*Acc.*
Video-LLaVA [4]	EMNLP 2024	1.96	35.8	2.02	34.9	2.94	57.5	2.24	42.8	2.29	42.7
LLaMA-VID [2]	ECCV 2024	2.09	36.6	2.07	36.0	2.83	56.9	2.23	42.6	2.30	43.1
MovieChat [7]	CVPR 2024	2.27	37.8	2.05	35.6	2.97	57.9	2.15	43.0	2.36	43.5
Video-online [11]	CVPR 2024	2.01	36.5	2.03	35.8	2.87	54.2	2.06	38.2	2.24	41.1
LongVA [20]	Arxiv 2024	2.48	47.1	2.74	45.4	2.98	57.8	2.22	42.4	2.60	48.1
LLaVA-Hound [29]	Arxiv 2024	2.69	48.7	2.56	43.7	3.07	56.8	**2.42**	42.7	2.68	47.9
FreeVA [8]	Arxiv 2024	2.48	46.7	2.32	41.7	3.02	58.1	2.16	38.3	2.49	46.2
Flash-VStream [10]	Arxiv 2024	2.02	37.3	2.06	36.1	2.91	56.1	2.08	39.8	2.26	42.3
STREAMCHAT	- -	**2.78**	**50.1**	**2.84**	**50.5**	**3.08**	**58.7**	2.38	**43.4**	**2.77**	**50.6**