WeaveTime: Stream from Earlier Frames into Emergent Memory in VideoLLMs

Supplementary Material

In Appendix, we provide additional information regarding:

- Additional Experimental Details in Sec. 1
- Additional Experimental Results in Sec. 2
- Qualitative Results in Sec. 3

1. Additional Experimental Details

In this section, we provide detailed explanations of the experimental setups corresponding to Fig. 2 and Tab. 1 in the main paper. We then present the full configuration of the SOPE training process, followed by Qwen-style input examples that illustrate how the model's input format differs during SOPE training and after training, compared with the original baseline input.

Ground-Truth Answer-Window Shift Details We conduct experiments on QAEGO4D [1], which provides explicit temporal answer windows suitable for controlled analysis. We randomly sample 100 videos and shift their answer windows to different temporal positions within the video (0%, 25%, 50%, 75%, and 100%), then record the model's predictions. Results show that for short videos, the model tends to focus on the beginning and middle–end regions, whereas for longer videos, it exhibits a strong bias toward the early portion of the video.

Human Evaluation with Shuffled-Frame Videos We further perform a human evaluation using VideoMME [5], a widely adopted video QA benchmark containing short, medium, and long videos, spanning diverse domains and twelve question categories. For each category, we uniformly sample videos and use ffmpeg [12] to shuffle their frames. As illustrated in Fig. 1, participants view the shuffled videos through our visualization interface and attempt to answer the provided questions. Even with shuffled frames, humans can often reconstruct the temporal flow using timestamps and correctly answer straightforward questions. However, tasks that require temporal reasoning, such as counting or action-based reasoning, become significantly more challenging when the temporal structure is disrupted. In contrast, Video-LLMs remain largely unaffected, indicating that current models treat videos as a bag of evidence rather than as inputs with a strict causal temporal order.

Training Configuration of SOPE. The Streaming Order Perception Enhancement (SOPE) is designed to strengthen a model's sensitivity to temporal structure and mitigate temporal order ambiguity. Table 1 summarizes the core training configurations. We adopt the transformers [15] training framework and fine-tune the model with LoRA [6] using only 30K sampled video instances, demonstrating that SOPE requires minimal data and computational overhead.

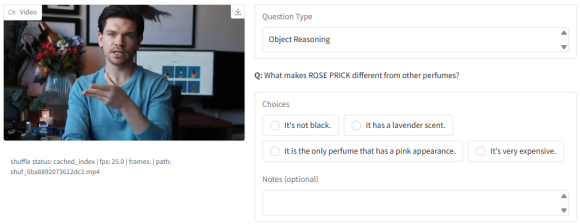


Figure 1. **Visualization Interface of Human Evaluation.** The interface presents the video frame on the left and the corresponding evaluation form on the right, enabling annotators to answer questions under shuffled-frame conditions and provide reliable human judgments for benchmarking.

		SOPE
Data	Dataset	LLaVA-Video-173K
	#Samples	30K
Model	Base Video-LLM	LLaVA-OV
	Trainable	Connector(full) LLM(LoRA r=128, scale=256)
Training	Batch Size per Device	1
	Gradient Accumulation	16
	Learning Rate	1e-5
	Warm-up Ratio	0.03
	LR Scheduler	Cosine
	Optimizer	Adamw
	Epochs	1
	Precision	bf16 fp16

Table 1. **Training Configuration of SOPE.**

Shuffle Input Examples. To illustrate how SOPE modifies the model's input format, we present Qwen-style prompt examples comparing the baseline and SOPE-enhanced inputs. For readability, we omit tokens such as "\n" and "vision-padd", and retain only "<im_start>" and "<im_end>" as delimiters. The placeholders `<frames = xxx s - xxx s>` denote grouped video-frame tokens covering the corresponding temporal intervals. Questions and answers are represented by generic `question` and `answer` placeholders.

Multi-turn Prompt

<im_start> system: You are a helpful assistant. <im_end>
<im_start> user: <frames=1.2s-3.9s> question 1? <im_end>
<im_start> assistant: answer 1. <im_end>
<im_start> user: question 2? <im_end>
<im_start> assistant: answer 2. <im_end>

Multi-turn Prompt (SOPE)

<im_start> system: You are a helpful assistant. <im_end>
<im_start> user:
t=0.3s-1.2s <frames=1.2s-2.1s>
t=1.2s-2.1s <frames=2.1s-3.0s>
t=2.1s-3.0s <frames=0.3s-1.2s>
t=3.0s-3.9s <frames=3.0s-3.9s>
These video segments are shuffled. List each segment's true time range. <im_end>
<im_start> assistant: Correct timestamps is t=1.2s-2.1s> | t=2.1s-3.0s | 0.3s-1.2s | 3.0s-3.9s. <im_end>
<im_start> user: question 1? <im_end>
<im_start> assistant: answer 1. <im_end>
<im_start> user: question 2? <im_end>
<im_start> assistant: answer 2. <im_end>

Shuffled-Video Re-order Prompts. The following list presents ten automatically generated questions produced using GPT. Each prompt requires the model to infer the correct temporal range or chronological order from disordered visual inputs. During training, one prompt is randomly sampled from this set for each instance.

Re-Order Questions

1. These video segments are shuffled. List each segment's true time range.
2. Segments have been randomly reordered. Output the correct timestamps (start–end) for each.
3. The segment order is randomized. Provide the real time span (start–end) for each segment.
4. Order and timestamps do not match. Write the true timestamps in chronological order.
5. Displayed timestamps may be wrong. Output the correct time range list.
6. Segments are shuffled. Infer and write each segment's correct timestamp.
7. Treat shown timestamps as distractors. Provide the real start–end times for each segment.
8. Recover the true temporal order and provide each segment's time range.
9. Current segment order is random. Output timestamps in actual occurrence order.
10. Time prompts do not match the videos. Provide each segment's correct time span.

2. Additional Experimental Results

As shown in Tab. 2, to demonstrate the generalizability of our WeaveTime approach, we conduct experiments on Qwen2-VL. The results show that our method consistently improves performance across diverse streaming scenarios and different baseline models. These findings indicate that effectively addressing time-agnosticism is crucial for robust streaming performance, and that our WeaveTime framework successfully mitigates this issue.

3. Qualitative Results

Fig. 2 presents additional qualitative results illustrating failure cases caused by the time-agnostic nature of existing models in streaming scenarios.

(a) Temporal Order Ambiguity. The baseline model struggles to interpret the video stream in its correct chronological order, leading to confusion between actions such as standing up and crouching. In contrast, Ours effectively resolves this ambiguity and correctly identifies the action as "She is crouching".

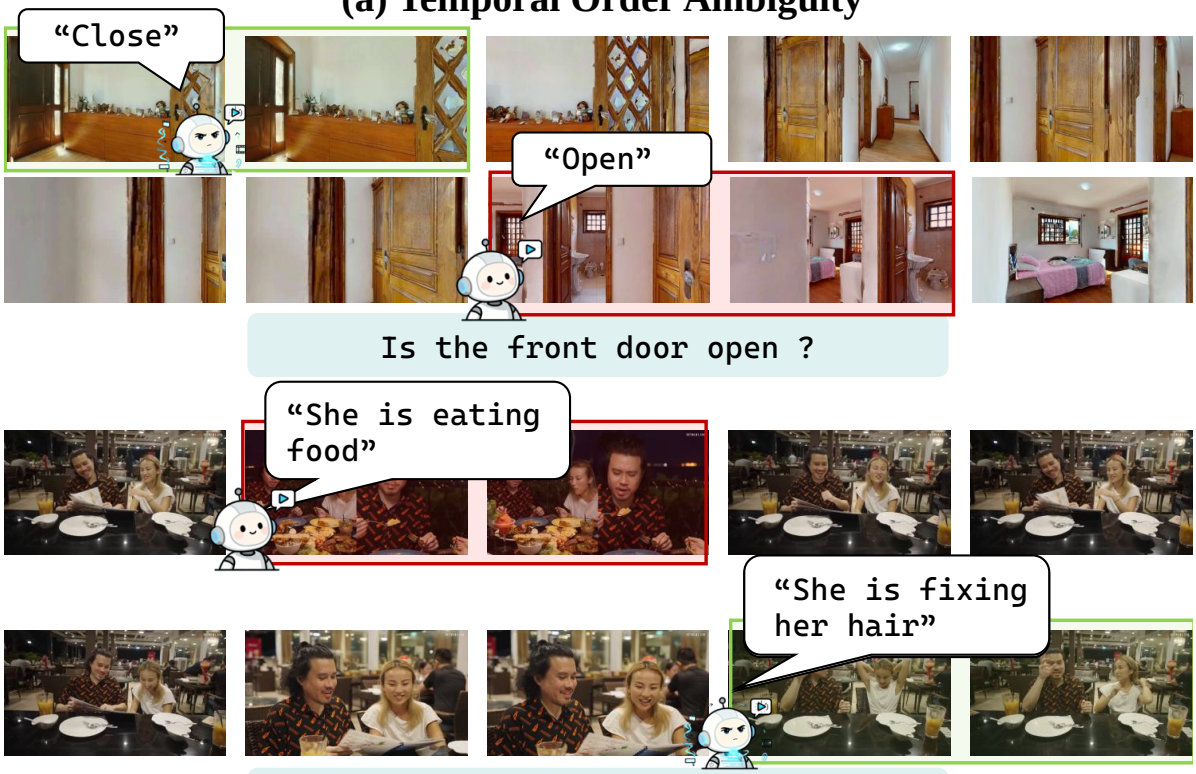
(b) Past–Current Focus Blindness. The baseline model fails to properly distinguish between past and present visual evidence. When asked about the state of the "front door", it relies solely on the current frame regardless of the temporal requirement; when queried about the ongoing action, it erroneously recalls an earlier eating moment. Ours accurately separates past from present cues and produces temporally grounded responses.

Method	# of Frames	OVO-Bench Real-Time							Streaming-Bench Real-Time										
		OCR	ACR	ATR	STU	FPD	OJR	AVG	OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	AVG
Human																			
Human	-	93.96	92.57	94.83	92.70	91.09	93.20	91.30	89.47	92.00	93.60	91.47	95.65	92.52	88.80	88.75	89.74	91.30	91.46
Proprietary Models (Offline), Single-Turn Evaluation																			
Gemini 1.5 pro [11]	1 FPS	85.91	66.97	79.31	58.43	63.37	61.96	69.32	79.02	80.47	83.54	79.67	80.00	84.74	77.78	64.23	71.95	48.70	75.69
GPT-4o [7]	64	69.80	64.22	71.55	51.12	70.30	59.78	64.46	77.11	80.47	83.91	76.47	70.19	83.80	66.67	62.19	69.12	49.22	73.28
Open-Source Models (Offline), Single-Turn Evaluation																			
Qwen2-VL-72B [14]	64	65.77	60.50	69.83	51.69	69.31	54.35	61.92	-	-	-	-	-	-	-	-	-	-	-
LLaVA-Video-7B [9]	64	69.13	58.72	68.83	49.44	74.26	59.78	63.52	-	-	-	-	-	-	-	-	-	-	-
LLaVA-OV-7B [8]	64/32	66.44	57.80	73.28	53.37	71.29	61.96	64.02	80.38	74.22	76.03	80.72	72.67	71.65	67.59	65.45	65.72	45.08	71.12
Qwen2-VL-7B [14]	64/1FPS	60.40	50.56	66.03	47.19	66.34	55.43	55.98	75.20	82.81	73.19	77.45	68.32	71.03	72.22	61.39	61.47	46.11	69.04
InternVL-V2-8B [3]	64/16	67.11	60.55	63.79	46.07	68.32	56.52	60.39	68.12	60.94	69.40	77.12	67.70	62.93	59.26	53.25	54.96	56.48	63.72
Open-Source Models (Streaming), Single-Turn Evaluation																			
Flash-VStream-7B [16]	1 FPS	24.16	29.36	28.45	33.71	25.74	28.80	28.37	43.59	25.87	24.91	23.87	27.33	13.08	18.52	25.20	23.87	48.70	23.23
VideoLLM-Online-8B [2]	2 FPS	8.05	23.85	12.07	14.04	45.54	21.20	20.79	39.07	40.06	34.49	31.05	45.96	32.48	31.40	43.16	42.49	27.89	35.99
VideoLLM-EyeWO-8B [17]	2 FPS	24.16	27.52	31.89	32.58	44.55	35.87	32.76	-	-	-	-	-	-	-	-	-	-	-
Dispider [10]	1 FPS	57.72	49.54	62.07	44.94	61.39	51.63	54.55	74.92	75.53	74.10	73.08	74.44	59.92	76.14	62.91	62.16	45.80	67.63
Open-Source Models (Offline → Streaming), Multi-Turn Evaluation																			
LLaVA-OV-7B																			
+ StreamBridge [13]	1 FPS	58.39	59.63	69.82	44.38	76.23	61.41	61.64	76.84	77.17	82.60	75.25	64.15	64.17	75.00	61.38	61.19	46.11	68.39
+ ReKV† [4]	1 FPS	63.09	55.05	72.41	46.63	72.28	60.87	61.72	65.85	78.12	77.92	71.84	66.88	65.11	69.44	62.20	61.08	43.09	66.15
+ WeaveTime (Ours)	1 FPS	72.48	69.72	74.13	53.37	75.24	67.93	68.82	71.54	81.25	86.75	78.64	75.16	73.21	72.22	69.11	68.75	44.68	72.13
Qwen2-VL-7B																			
+ StreamBridge [13]	1 FPS	65.10	64.22	64.66	46.63	74.26	65.22	63.35	80.38	78.74	83.22	79.86	74.21	69.47	77.78	63.41	69.97	43.01	72.01
+ ReKV† [4]	1 FPS	60.40	52.29	68.10	43.26	72.28	61.96	59.72	71.00	82.03	80.13	74.76	71.34	69.16	73.15	64.63	68.75	45.74	70.07
+ WeaveTime (Ours)	1 FPS	75.17	59.63	71.55	51.69	72.28	67.39	66.28	74.80	85.94	88.01	81.23	77.71	76.32	82.41	63.82	74.15	49.47	75.39

Table 2. Comparison of various Video LLMs on **OVO-Bench Real-Time** and **Streaming-Bench Real-Time**. † indicates experimental results derived from integrating the respective dataset into the ReKV[4] codebase.

What is her posture while cleaning the mirror ?

(a) Temporal Order Ambiguity



Is the front door open ?

What is she doing ?

(b) Past–Current Focus blindness

Figure 2. **Qualitative Results.**

References

[1] Leonard Bärmann and Alex Waibel. Where did i leave my keys? — episodic-memory-based question answering on egocentric videos. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1559–1567, 2024. ISSN: 2160-7516. 1

[2] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video, 2024. 3

[3] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, Ji Ma, Jiaqi Wang, Xiaoyi Dong, Hang Yan, Hewei Guo, Conghui He, Botian Shi, Zhenjiang Jin, Chao Xu, Bin Wang, Xingjian Wei, Wei Li, Wenjian Zhang, Bo Zhang, Pinlong Cai, Licheng Wen, Xiangchao Yan, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites, 2024. 3

[4] Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, Hao Jiang, et al. Streaming video question-answering with in-context video kv-cache retrieval. In *ICLR*, 2025. 3

[5] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis, 2025. 1

[6] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models, 2021. 1

[7] Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, et al. Gpt-4o system card, 2024. 3

[8] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer, 2024. 3

[9] Feng Li, Renrui Zhang, Hao Zhang, Yuanhan Zhang, Bo Li, Wei Li, Zejun Ma, and Chunyuan Li. Llava-next-interleave: Tackling multi-image, video, and 3d in large multimodal models, 2024. 3

[10] Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction, 2025. 3

[11] Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, Soroosh Mariooryad, Yifan Ding, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context, 2024. 3

[12] Suramya Tomar. Converting video formats with ffmpeg. *Linux Journal*, 2006(146):10, 2006. 1

[13] Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. Streambridge: Turning your offline video large language model into a proactive streaming assistant, 2025. 3

[14] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution, 2024. 3

[15] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Huggingface's transformers: State-of-the-art natural language processing, 2020. 1

[16] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams, 2024. 3

[17] Yulin Zhang, Cheng Shi, Yang Wang, and Sibei Yang. Eyes wide open: Ego proactive video-llm for streaming video, 2025. 3