STREAMMIND: Unlocking Full Frame Rate Streaming Video Dialogue through Event-Gated Cognition

Supplementary Material

1. Impact of Silence-Response Sample Imbalance

As described in the data preparation process, during the Silence-Response Labeling stage, the dataset inherently exhibits a severe imbalance in the Silence-Response ratio. For example, in a 1-minute video at 30 FPS, if only 5 frames require a response, the silence-response ratio becomes 360:1. Such an extreme imbalance can significantly impact the training process.

To mitigate this issue, we introduce a balancing weight W_s into the standard Cross Entropy (CE) loss during training, where W_s and $1 - W_s$ represent the weights for silence and response, respectively. We conducted extensive ablation studies on Silence-Response sample balancing based on the weight W_s. The results indicate that, compared to the standard cross-entropy loss, W_s significantly improves the final performance. However, an unfortunate observation is that the optimal W_s varies considerably, even differing by orders of magnitude. To address this, we analyzed the ratio P of $< /silence >$ to $< /response >$ tokens in the Ego4D and SoccerNet datasets, which are $310 : 1$ and $71 : 1$, respectively. Based on these observations, we derive an empirical formula for the optimal balancing weight W_s^{opt}:

$$W_s^{\text{opt}} \approx 10P \tag{1}$$

Table 1. Perceptive Judgement Capability

Datasets	Method	TimeDiff ↓	TriggerAcc ↑	TimVal ↑
Ego4D	Standard CE	2.05	31.37%	30.91%
	0.10	1.92	41.34%	37.60%
	0.12	1.88	42.16%	38.63%
	0.15	1.89	43.34%	39.73%
	0.17	1.95	40.23%	39.44%
	0.20	1.97	35.35%	34.60%
Soccer	Standard CE	15.89	31.34%	29.61%
	0.01	14.33	50.35%	41.32%
	0.02	14.35	51.62%	44.8%
	0.03	14.02	52.18%	47.36%
	0.04	14.03	44.69%	42.37%
	0.05	14.12	41.22%	39.82%

2. Perception Phase Visualization Experiment

To visualize the perception phase, we conducted the following experiment: As streaming video inputs continued, we preserved all perception tokens generated by the Cognition Gate between two consecutive cognition phases. We then computed the cosine similarity between these tokens to analyze their consistency over time.

The results indicate that perception tokens generated by EPFE can effectively distinguish between relevant and irrelevant events while maintaining a strong memory of the primary event. Even after encountering unrelated events, the perception tokens are capable of refocusing on the main event. Furthermore, throughout the entire event, perception tokens exhibit robust memory retention, maintaining a high similarity with earlier stages of the event even in later phases, as shown in figure.1.(a).

We repeated this experiment on STC, as shown in figure 1.(b), but its perception tokens failed to maintain long-term feature similarity, capturing only local spatiotemporal features. This further highlights the superiority of EPFE in preserving event-level semantics over extended video sequences.

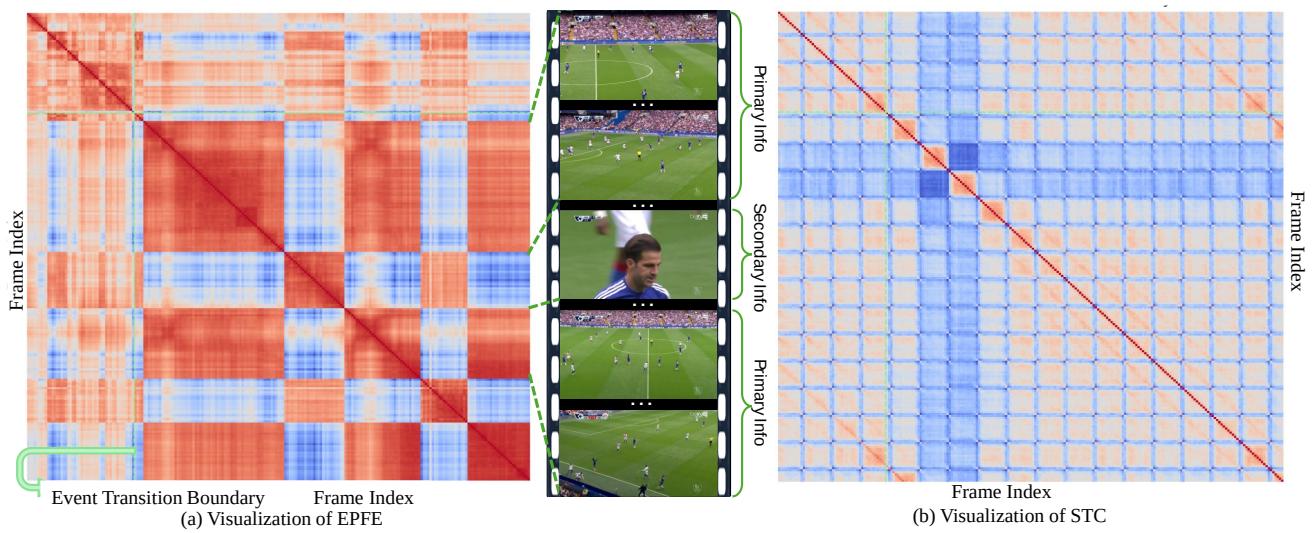
3. A Discussion on Event Perception Mechanisms and Model Design

In StreamMind's implementation, an 'event' refers to "a series of frames within a time segment, after which the Cognition Gate determines that sufficient information relevant to the user query has been gathered and forwards it to the LLM for cognition". It aligns with works [3], defines events as "A segment of time... have a beginning and an end". As noted in work [4], while the human sensory system receives data at 1e9 bits/s, it processes only 10 bits/s—highlighting the Gate's role in efficiently filtering vast sensory input for cognition.

4. Extended Benchmark Evaluation

To further demonstrate the superior performance of StreamMind, we conducted additional comparisons with FlVStream, Video-Online [1], and Dispider [2] on OVO-Bench and StreamingBench, as shown in Table 3. It can be observed that, while StreamingBench is not specifically designed for the Streaming Video Dialogue (SVD) task—focusing mainly on understanding rather than timing alignment and real-time performance (except for PO)—our method nonetheless matches or outperforms the baselines in these metrics, while maintaining the speed advantages we emphasize. In contrast, we achieve significantly better performance on OVO-Bench (7% improvement over Dispider), a benchmark explicitly designed for the SVD task.

5. Visualization of demo



(a) Visualization of EPFE (b) Visualization of STC

Figure 1. A cosine similarity heatmap of frames across two consecutive events.

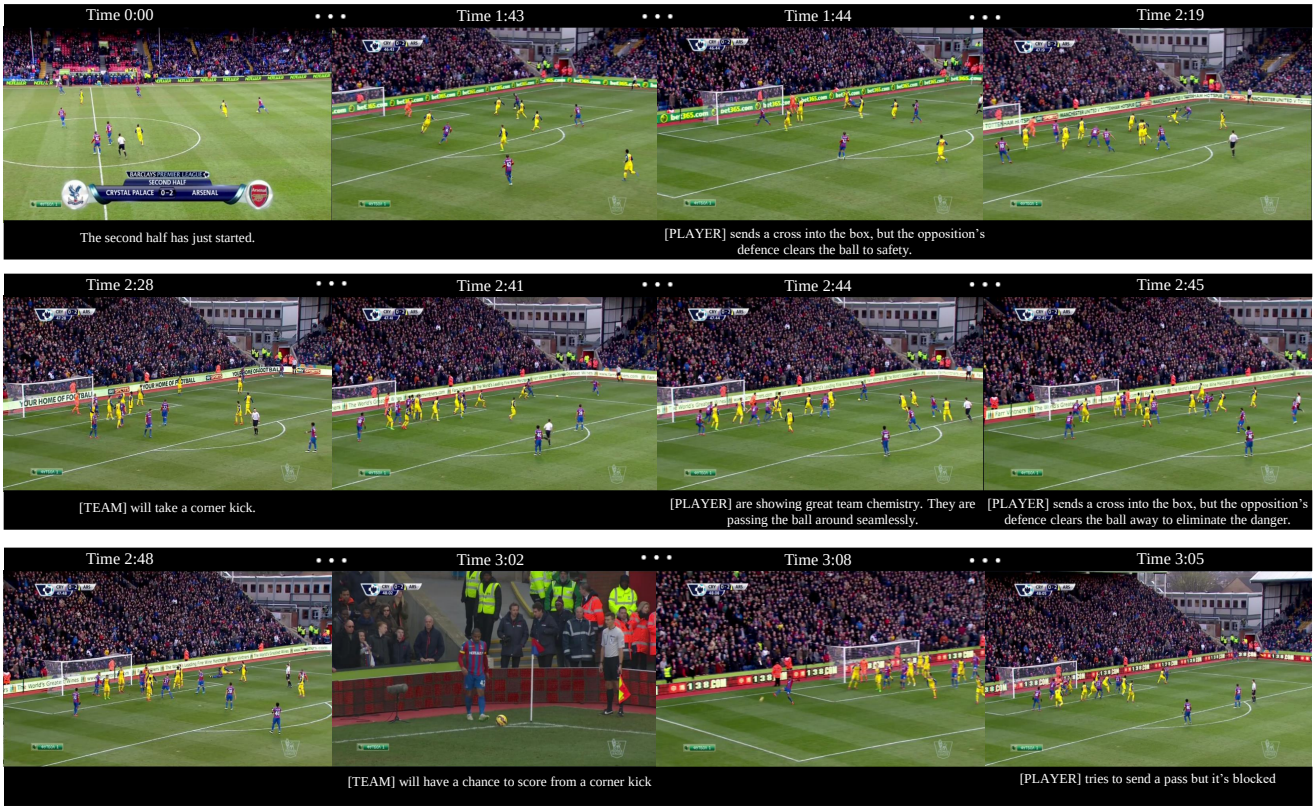


Figure 2. An illustrative snippet of StreamingVD on a live football match. The query at the beginning is: "Hey, Robot, can you watch the football game with me and provide commentary?"

References

[1] 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. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18407–18418, 2024. 1

[2] Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider:

Table 2. **Performance comparison on OVO-Bench with StreamingVLM**

Model	Frames	OVO-Bench												
		Real-Time Perception						Backward			Forward Active			Avg.
		OCR	ACR	ATR	STU	FPD	OJR	EPM	ASI	HLD	REC	SSR	CRR	
Human	-	93.9	92.6	94.8	92.7	91.1	94.0	92.6	93.0	91.4	95.5	89.7	93.6	92.8
Fl-VStream	1fps	24.2	29.7	28.5	33.7	25.7	28.8	39.1	37.2	5.91	8.0	67.3	60.0	33.6
Video-online	2fps	8.1	23.9	12.1	14.0	45.5	21.2	22.2	18.8	12.18	-	-	-	-
Dispider	1fps	57.7	49.5	62.1	44.9	61.4	51.6	48.5	55.4	4.3	18.1	37.4	48.8	41.8
Streammind	1fps	58.9	50.7	62.3	45.3	61.1	51.3	47.6	55.4	15.8	19.9	42.6	51.6	46.9
Streammind	2fps	59.7	52.4	63.3	45.4	61.8	51.4	48.5	54.8	17.9	25.7	44.5	52.1	48.2

Table 3. **Performance comparison on StreamingBench with StreamVLM**

Model	Frames	StreamingBench																	
		Real-Time understanding										Omni-source				Contextual				All
		OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	ER	SCU	SD	MA	ACU	MCU	SQA	PO	
Human	-	89.5	92.0	93.6	91.5	95.6	92.5	88.0	88.8	89.7	91.3	88.0	88.2	93.6	90.3	88.8	90.4	95.0	100	91.7
Fl-VStream	1fps	25.9	43.6	24.9	23.9	27.3	13.1	18.5	25.2	23.9	48.7	25.9	24.9	25.6	28.4	24.8	25.2	26.8	1.96	24.0
Video-online	2fps	39.1	40.1	34.5	31.1	45.9	32.4	31.5	34.2	42.5	27.9	31.2	26.5	24.1	32.0	24.2	29.2	30.8	3.92	32.5
Dispider	1fps	74.9	75.5	74.1	73.1	74.4	59.9	76.1	62.9	62.2	46.7	35.5	25.3	38.6	43.3	39.6	27.7	34.8	25.3	53.1
Streammind	1fps	74.8	75.8	75.2	73.8	73.9	61.3	75.5	64.2	62.4	47.1	36.2	25.6	34.7	43.7	42.5	30.3	36.8	37.5	54.0
Streammind	2fps	75.4	76.4	75.5	73.2	74.6	61.2	76.3	64.5	62.3	47.8	36.5	25.6	40.4	45.3	44.6	31.4	35.9	39.8	55.2

Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24045–24055, 2025. 1

[3] Jeffrey M Zacks, Nicole K Speer, Khena M Swallow, Todd S Braver, and Jeremy R Reynolds. Event perception: a mind-brain perspective. *Psychological bulletin*, 133(2):273, 2007. 1

[4] Jieyu Zheng and Markus Meister. The unbearable slowness of being: Why do we live at 10 bits/s? *Neuron*, 113(2):192–204, 2025. 1