TOWARDS TRAINING-FREE PROACTIVE RESPONSES FOR VIDEO MULTIMODAL LARGE LANGUAGE MODELS VIA CUMULATIVE INFORMATION ESTIMATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite great success, existing *multimodal large language models* (MLLMs) still look *"dumb"* in online video interaction, primarily due to the lack of *proactive response* capability. In this paper, we propose a novel and plug-and-play approach for MLLMs, termed *Cumulative Information based Active response* (**CIAct**). Unlike existing methods that typically frame proactive response as a multi-class prediction of model states, CIAct formulates it as a problem of *cumulative information estimation*, *i.e.*, allowing the MLLM to respond to questions once relevant information is accumulated enough. This definition not only allows CIAct to leverage existing lightweight vision-language models to help perceive query-related information, but also avoids the expensive MLLM-based video watching and dedicated SFT tuning. Moreover, CIAct is also equipped with a carefully designed memory replay scheme, helping it to cope with different online QA tasks efficiently, such as *backward tracing*, *real-time perception*, and *forward active responding*. To validate CIAct, we apply it to two representative offline MLLMs, and compare to a set of proactive response models and methods on both online and offline benchmarks. The experimental results not only show its superior performance than existing methods on streaming benchmarks, e.g., +4.04% on OVO-Bench, but also confirm its great benefits to existing offline MLLMs. For instance, CIAct helps Qwen3-VL save about 38% of watching time on LVBench, while obtaining a +2.4% performance gain. **Our code** is in the supplementary materials.

1 INTRODUCTION

For a year or two, great breakthroughs have been made by a set of *multimodal large language models* (MLLMs) in a variety of image or video tasks (Liu et al., 2023; Li et al., 2024; Yang et al., 2025a; Zhang et al., 2024b; Chen et al., 2024c; 2025; Tong et al., 2025). Although exhibiting strong capabilities in video understanding, existing MLLMs struggle to handle the major scenarios of online video interaction, *i.e.*, answering questions based on continuously arriving video streams. This is mainly because that most MLLMs are designed for offline video tasks (Maaz et al., 2024; Zhang et al., 2023; Li et al., 2025; Lin et al., 2024) and formulate video understanding as a multi-image based QA task (Li et al., 2024; Zhang et al., 2024b; Wang et al., 2024; Bai et al., 2025; Chen et al., 2024b), thus requiring the complete input for instant inference. In this case, most MLLMs are in absence of *proactive response* capability (Chen et al., 2024a), *i.e.*, knowing when and what to answer given the streaming videos, which however is critical for online video applications.

The difficulty of cultivating proactive response for existing MLLMs (Chen et al., 2024a; Qian et al., 2025; Wang et al., 2026a;b; Zhang et al., 2026a; Yan et al.) lies in two main aspects. Firstly, under online (streaming) settings, proactive response requires MLLMs to reason from incomplete observations and respond to users once answer cues are sufficient (Zhang et al., 2026b). But in practice, MLLMs are often stuck in a dilemma of instant decision with insufficient information (Tong et al., 2026) or prolonged observations with drastically increasing latency (Wang et al., 2026b). Secondly, online questions can arrive at any time, but the answer cues may already emerge in past video content (Niu et al., 2025). This also raises a new concern that *knowing when to answer* (Kim et al., 2025) is already not enough for online MLLMs.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

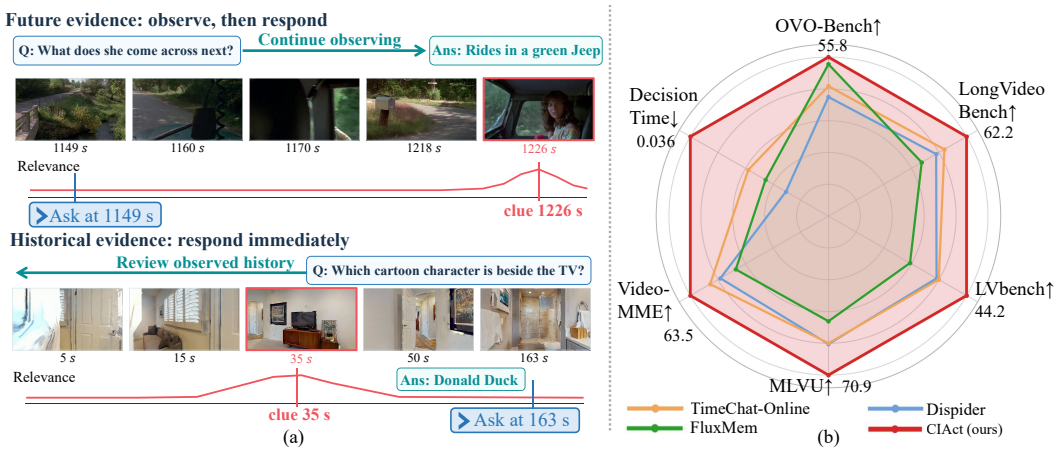


Figure 1: (a) Illustration of the proposed CIAct. As a plug-and-play method, CIAct can help existing MLLMs to accomplish proactive responses in terms of *backward tracking*, *real-time* or *forward answering* in the online video scenario. (b) Comparison with existing approaches. CIAct not only achieves better performance than existing proactive response models on both streaming and offline[1] benchmarks, but also exhibits superior efficiency and requires no additional training.

To realize proactive response for MLLMs, one popular solution is to build dedicated datasets for SFT tuning. For instance, Streamo (Xia et al., 2026) and VideoChat3 (Li et al., 2026) carefully collect a large number of examples for the online-setting tuning of MLLMs. Except for the expensive expenditure of annotation and SFT tuning, this paradigm inevitably encounters the issue of increasing latency and overhead caused by the accumulated video frames under the online setting (Chen et al., 2024a; Xia et al., 2026). For instance, Streamo uses the full-scale MLLM at each timestamp to predict model states based on accumulated video frames. There are also some attempts (Xie et al., 2026; Zhang et al., 2026a) conducted for training-free proactive response, which use a lightweight *vision-language* (VL) model (Radford et al., 2021) to trigger a response when substantial scene changes are detected. However, this content-change based solution is hard to well handle the diverse online video interactions (Zhang et al., 2026a), *e.g.,* the backward tracking one. In this case, how to realize effective and efficient online-MLLMs still remains an open problem.

In this paper, we propose a novel and training-free approach for the proactive response of MLLMs, termed *Cumulative-Information based Active response* (**CIAct**). Unlike existing methods that typically define proactive response as a multi-class prediction task, CIAct formulates this target as a problem of *cumulative information estimation*, *i.e.*, deciding when the received information is sufficient for answering by MLLMs. This reformulation can transform the costly online video modeling into independent steps of relevant information measurement and cumulative state estimation, thereby greatly saving the expenditure of proactive response. In practice, CIAct calibrates frame-query relevance by lightweight VL models (Li et al., 2022; Tschannen et al., 2025), and accumulates it with a decayed *cumulative sum* (CUSUM) function (Page, 1954) for state estimation. Moreover, CIAct is also equipped with a memory replay scheme to integrate answer ques appearing before and after the query, helping it better handle the diverse tasks of online video interaction. The overall process requires no training and can be plugged into existing MLLMs.

To validate CIAct, we apply it to LLaVA-OV-7B (Li et al., 2024) and Qwen3-VL-8B (Yang et al., 2025a) and conduct experiments on a set of streaming and offline video benchmarks (Wu et al., 2024; Wang et al., 2025a; Fu et al., 2025; Zhou et al., 2025). The experimental results not only demonstrate the superior performance and efficiency of CIAct over specifically tuned MLLMs and other methods, *e.g.*, a +4.04% improvement on OVO-Bench, but also confirm its great benefits for existing offline MLLMs. For instance, under the proactive response setting, CIAct helps Qwen3-VL save 38% of watching time and improve accuracy by 2.4% on LVBench.

Overall, our main contributions are threefold:

[1]Evaluated using the proactive response protocol.

- We formulate the proactive response of MLLMs as a problem of cumulative information estimation, and propose a novel and training-free approach termed CIAct.
- The proposed CIAct can apply lightweight VL models to estimate information accumulation without specific annotation tuning, and adopts relevance-driven memory replay for fast online interaction, greatly reducing the cost of proactive response.
- The proposed CIAct not only yields new SOTA performance on streaming benchmarks, but also shows great benefit for existing MLLMs in offline video understanding.

2 Related Work

Video Multimodal Large Language Models. Video multimodal large language models (Video-MLLMs) have evolved from video-specific extensions of image MLLMs to general-purpose models that jointly support images and videos. Their main principle is to convert sampled video frames into visual tokens and use them to condition an autoregressive language model. Most Video-MLLMs follow a common architecture consisting of a visual encoder, a temporal aggregation module or position encoding, a multimodal projector, and a large language model. Early methods, such as Video-ChatGPT (Maaz et al., 2024) and Video-LLaVA (Lin et al., 2024), adapt pretrained image encoders to video inputs, while recent models, including LLaVA-OneVision (Li et al., 2024), Qwen2-VL (Wang et al., 2024), and VideoChat3 (Li et al., 2026), adopt unified visual representations and more flexible video tokenization. Their training generally follows two stages: vision-language alignment on image/video-text pairs, followed by multimodal instruction tuning on captioning, question-answering, and conversational data (Lin et al., 2024; Li et al., 2024; Wang et al., 2024). Despite their strong offline video understanding capabilities, these models usually assume that the complete video is available before inference. They sample and encode a fixed set of frames as a batch and generate an answer only after receiving the full visual context. Consequently, they cannot naturally update their state over an unbounded stream or autonomously determine when the observed evidence is sufficient for a response. Our work addresses this limitation by enabling existing Video-MLLMs to make proactive responses through training-free evidence accumulation.

Proactive Responses in Online Video Streams. Existing approaches to proactive response can be broadly divided into streaming-MLLM, agent-based, and scene-based methods. Streaming-MLLM methods learn when to answer by predicting special tokens or discrete response states from online interaction data (Chen et al., 2024a; Xia et al., 2026; Li et al., 2026; Wang et al., 2025b; 2026b; Yan et al.). However, they require additional annotation and tuning, while repeatedly processing accumulated video frames with an MLLM introduces increasing latency. Agent-based methods decouple response decisions from answer generation and use a lightweight model to decide whether to wait or answer at each timestamp (Wang et al., 2026a; Qian et al., 2025). This design reduces decision costs but still relies on an additionally trained module for instantaneous state prediction. Scene-based methods provide a training-free alternative by triggering responses according to visual changes, query relevance, or information novelty (Yao et al., 2025; Xie et al., 2026; Zhang et al., 2026a; Zheng et al., 2026). However, visual changes do not necessarily provide query-relevant information, making such rules difficult to handle diverse online interactions. In contrast, CIAct formulates proactive response as cumulative information estimation. It accumulates background-calibrated query relevance with a decayed CUSUM and responds once the received information is sufficient for answering.

3 Preliminary

Given a video stream V and a question Q, *proactive response* requires the MLLM to determine when and what to answer based on the observed content $V_{1:t}$, of which objective can be defined by

$$(\hat{r}_t, \hat{y}_t) = \arg\max_{r,y} \underbrace{P(r \mid V_{1:t}, Q)}_{\text{Response decision}} \underbrace{P(y \mid r, V_{1:t}, Q)}_{\text{Answer generation}}, \tag{1}$$

where r_t denotes the response decision at the t-th time step, and y is the answer generated based on the observed video content $V_{1:t}$ for the question Q.

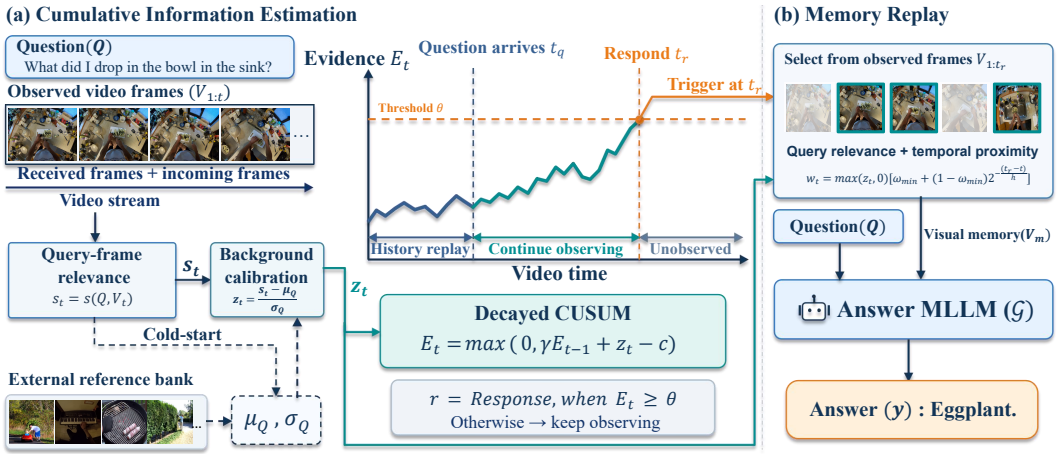


Figure 2: Overview of CIAct. CIAct helps existing MLLMs achieve proactive response via *Cumulative Information Estimation* (a). Given the video stream, when a query arises, CIAct will use the CUSUM function to estimate the response state based on query-frame relevances of received frames. When the cumulative state reaches the threshold, it suggests that relevant information is sufficient, so the response of MLLMs can be triggered. Otherwise, CIAct will continuously observe the upcoming frames. When responding, CIAct will conduct *Memory Replay* (b) to select suitable video frames for MLLMs, thereby maintaining budgeted computation overhead. CIAct is a training-free approach that enables the plug-and-play ability of proactive response for existing MLLMs.

For the above objective, a direct solution is to let the MLLM $\mathcal{G}(\cdot)$ be aware of the right moment to respond, and existing methods (Qian et al., 2025; Xia et al., 2026; Li et al., 2026) typically formulate it as a multi-class prediction of model states:

$$r_t = \mathcal{G}(V_{1:t}, Q), \qquad \text{where } r_t \in \{\text{Silence}, \text{Standby}, \text{Response}\}. \tag{2}$$

Here, *Silence*, *Standby* and *Response* are three states of *no relevant evidence*, *insufficient information* and *sufficient evidence* for answering, respectively.

However, this definition requires dedicated training of MLLMs on carefully collected video examples with labor-intensive annotations. Furthermore, even with time-window constraints (Ding et al., 2025; Kim et al., 2026), the long video streams also impose significant computation and memory overhead by the continuous processing of video frames in MLLMs.

There are also attempts conducted for the training-free active response for MLLMs (Xie et al., 2026; Zhang et al., 2026a). And their main principle is to trigger the response by perceiving drastic video content changes, *i.e.*, letting the model answer *when the current video event concludes*:

$$r_t = \begin{cases} Response, & 1 - \cos(V_{1:t-1}, V_t) \geq \theta, \\ Wait, & \text{otherwise.} \end{cases} \tag{3}$$

Here, $\cos(V_{1:t-1}, V_t)$ denotes the visual similarity between the historical observations and the current frame, and θ is the trigger threshold. Although much more efficient, this solution struggles to accommodate the complex scenarios found in online interactions, *e.g.*, cases where the model is required to provide an immediate response or review historical video frames.

In this paper, we aim to study the training-free proactive response of MLLMs from the perspective of *cumulative information estimation*, targeting efficient and robust online video interactions.

4 Method

In this paper, we propose a novel and plug-and-play approach for existing MLLMs towards efficient and effective proactive response, termed *Cumulative-Information based Active response* (**CIAct**).

Different from existing methods, CIAct formulates proactive response as a task of *cumulative information estimation*, *i.e.*, estimating whether sufficient query-related information is perceived for answering. Thus, the objective of Eq. 1 can be re-formulated by

$$(\hat{r}_t, \hat{y}_t) = \arg\max_{r,y} \underbrace{P(r \mid \mathcal{E}(V_{1:t}, Q; \theta))}_{\text{Response decision}} \underbrace{P(y \mid r, \mathcal{G}(V_{1:t}, Q))}_{\text{Answer generation}}, \tag{4}$$

where $\mathcal{E}(\cdot)$ denotes the cumulative estimation function, θ is the response threshold, and $\mathcal{G}(\cdot)$ denotes the MLLM, while r and y represent the response decision and generated answer, respectively.

To achieve the above objective, we adopt the *cumulative sum* (CUSUM) function (Page, 1954) to model the query-relevant cumulative information state in video streams. Specifically, CIAct updates its cumulative information state E_t at each timestamp by

$$E_t = \max\left(0, \gamma E_{t-1} + z_t - c\right), \quad \text{with} \quad E_0 = 0. \tag{5}$$

Here, $\gamma \in (0, 1]$ is a *temporal decay* factor, and $c \geq 0$ is a *reference offset* that suppresses small deviations caused by the noisy VL alignment. And z_t is the calibrated query–frame relevance, which is defined by

$$z_t = \frac{s_t - \mu_Q}{\sigma_Q}, \tag{6}$$

where s_t denotes the raw relevance between question Q and frame V_t, *e.g.*, feature similarity, while μ_Q and σ_Q denote the query-specific *background mean* and *standard deviation*, respectively.

In principle, Eq. 5 treats query-relevant information as a positive deviation from the main content of a video stream. Observations with $z_t > c$ contribute positive increments, allowing individually weak but persistent cues to be accumulated for proactive response.

Based on the cumulative information state, CIAct determines the proactive response once the cumulative state reaches the threshold:

$$r = \mathit{Response}, \quad \text{when} \quad E_t \geq \theta. \tag{7}$$

For the second term of Eq. 4, *i.e.*, $P(y \mid r, \mathcal{G}(V_{1:t}, Q))$, inputting all observed video frames $V_{1:t}$ into the MLLM will incur rapidly increasing computational and memory costs as the video proceeds. Restricting the input to a fixed-length temporal window, however, may discard distant historical evidence required for *backward tracing* tasks (Niu et al., 2025), as discussed above.

To address this issue, we further introduce a *memory replay* setting that reduces the response's cost while retaining relevant video frames. Accordingly, the objective in Eq. 4 is reformulated as

$$(\hat{r}_t, \hat{y}_t) = \arg\max_{r,y} \underbrace{P(r \mid \mathcal{E}(V_{1:t}, Q; \theta))}_{\text{Response decision}} \underbrace{P(y \mid r, \mathcal{G}(V_m, Q))}_{\text{Answer generation}}, \tag{8}$$

where V_m denotes the compact set of video frames selected by CIAct from the observed video.

Overall, CIAct enables efficient and effective proactive response for existing MLLMs without labor-intensive annotations and expensive model tuning.

Cold start of CIAct. The applied CUSUM function relies on a well-defined background distribution (Page, 1954). Specifically, the background mean μ_Q and standard deviation σ_Q in Eq. 6 provide the reference statistics for quantifying the deviation of each observation. In conventional settings, these statistics are commonly estimated from historical data before online detection.

However, for the proactive response of MLLMs, a question may arrive at the beginning of the online video, or the answer cues can appears before the question arrives. In this case, CIAct involves the *cold-start* problem for its cumulative information estimation.

To address this issue, we use an external reference frame bank $\mathcal{V}_{\text{ref}} = \{V_{\text{ref}}^1, \ldots, V_{\text{ref}}^M\}$, independent of the current video, and estimate μ_Q and σ_Q by

$$\mu_Q = \frac{1}{M} \sum_{j=1}^{M} s(Q, V_{\text{ref}}^j), \qquad \sigma_Q = \sqrt{\frac{1}{M} \sum_{j=1}^{M} \Big(s(Q, V_{\text{ref}}^j) - \mu_Q\Big)^2}. \tag{9}$$

Here, M is the number of reference frames. Via Eq. 9, CIAct requires no waiting for background samples from the current stream.

Visual Memory Recall To reduce the length of input video frames and keep a proper and relevant coverage of past information, we further introduce a memory-replay design for CIAct.

Let t_r denote the time step at which the model decides to respond. For each observed frame V_t, where $t \leq t_r$, we compute its memory weight by jointly considering its query relevance and temporal proximity to the response time via a *time-decay weighting* function (Ding & Li, 2005):

$$w_t = \max(z_t, 0) \left[\omega_{\min} + (1 - \omega_{\min}) 2^{-(t_r - t)/h}\right], \tag{10}$$

where $h > 0$ denotes the temporal *half-life* that controls how quickly the weight decreases with the distance from the response time t_r, and $\omega_{\min} \in [0, 1]$ specifies the minimum temporal weight. Consequently, the temporal weight equals 1 at t_r and gradually approaches $\omega_{\min}$ for earlier frames.

Based on Eq. 10, we can select up to K frames with the highest weights and arrange them in temporal order to form the visual memory V_m:

$$V_m = \mathrm{TopK}_{V_t \in V_{1:t_r}}(w_t). \tag{11}$$

This allows the model to make effective use of the observed evidence while reducing the computational cost of answer generation.

Cases of online video interaction. In practice, the applications of CIAct can be categorized into two main types. The first one is the commonly used *streaming* setting (Niu et al., 2025), where question can arrive at any time step and the tasks include *backward*, *real-time* and *forward* question answering. For this setting, considering the question arrives at the t-th time step, CIAct first reviews all historical video frames and obtains the current cumulative state E_t. If E_t is large enough for response, CIAct will have MLLMs to immediately answer the question with the selected memory V_m. Otherwise, CIAct will continue to watch videos and update E_t.

The other one is *watching video with questions*. Under this setting, the question will be given at the beginning of the video, *i.e.*, the 0-th step, and CIAct continuously evaluates the in-coming video frames under online settings until sufficient information is accumulated for response. This setting is also the one that our paper advocates, which can reduce the overhead of common offline video understanding under online settings. *i.e.*, not need to wach the entire video for question answering. The detailed implementation of our CIAct can refer to Sec. 5.1 and the Appendix.

5 Experiments

5.1 Experimental Settings

Benchmarks and Metrics. We compare CIAct with existing methods on the streaming benchmark OVO-Bench (Niu et al., 2025) and four offline benchmarks: LongVideoBench (Wu et al., 2024), LVBench (Wang et al., 2025a), Video-MME (Fu et al., 2025), and MLVU (Zhou et al., 2025). OVO-Bench includes three task categories, namely, *backward tracing*, *real-time perception*, and *forward active response*. We report the score of each category and the overall score. For offline benchmarks, in addition to accuracy, we also report the *average fraction of video viewed* (Avg. Viewed).

Compared Methods. The compared methods are divided primarily into two categories. The first one is the training-based online MLLMs, including Flash-VStream (Zhang et al., 2024a), VideoLLM-online (Chen et al., 2024a), Dispider (Qian et al., 2025), TimeChat-Online (Yao et al., 2025), and Streamo (Xia et al., 2026). The other one is the training-free approaches, such as StreamAgent (Yang et al., 2025b), QueryStream (Zhang et al., 2026a), and FluxMem (Xie et al., 2026).

Evaluation Protocols. The evaluation protocols of the compared methods fall into three types. The first one is the **offline** protocol, *i.e.*, giving the entire streaming video for answering, which is mainly used for offline MLLMs such as Qwen3-VL (Yang et al., 2025a). The second one is what we call the **semi-online** protocol. This protocol will provide the time for response, and the model only needs to consider what to answer given the observed videos, which are widely used for methods focusing on streaming video compression (Chen et al., 2026; Song et al., 2026). The last one is the most rigorous **proactive response** protocol, where both response time and answer content are decided by the model itself, *e.g.*, Streamo (Xia et al., 2026). In this paper, all experiments of our CIAct follow the proactive response protocol, also including tests on offline benchmark.

Table 1: Performance comparison with existing methods on OVO-Bench. *Frames* denotes the video sampling frequency, and *Input* denotes the maximum number of frames provided to the answering MLLM. ○, □, and △ denote the *offline*, *semi-online*, and *proactive response* protocols, respectively.

Models	Frames	Input	Backward Tracing				Real-Time Visual Perception							Forward Act. Responding				Overall
			EPM	HLD	ASI	Avg.	ACR	ATR	OJR	STU	OCR	FPD	Avg.	REC	CRR	SSR	Avg.	Avg.
Training-based Proactive Response MLLMs																		
Flash-VStream-7B □	1 FPS	1 FPS	39.06	5.91	37.16	27.38	29.36	28.45	28.80	33.71	24.16	25.74	28.37	8.02	60.00	**67.25**	45.09	33.61
VideoLLM-online-8B □	2 FPS	2 FPS	22.22	12.18	18.80	17.73	23.85	12.07	21.20	14.04	8.05	45.54	20.79	–	–	–	–	–
Dispider-7B □	1 FPS	1 FPS	48.48	4.30	55.41	36.06	49.54	62.07	51.63	44.94	57.72	61.39	54.55	18.05	48.75	37.36	34.72	41.78
TimeChat-Online-7B □	1 FPS	1 FPS	53.90	9.10	62.80	42.00	48.60	64.70	55.40	44.90	69.80	68.30	58.60	32.50	40.00	36.50	36.40	45.60
Streamo-7B △	1 FPS	1 FPS	54.55	**31.72**	52.03	46.10	57.80	75.00	**70.11**	49.44	79.19	64.36	65.98	29.96	**83.33**	51.03	**54.77**	55.61
Training-free Proactive Response MLLMs																		
StreamAgent-7B △	1 FPS	1 FPS	54.80	25.80	58.10	41.70	53.20	63.60	58.70	53.90	71.20	67.30	61.30	35.90	52.00	48.40	45.40	49.40
QueryStream-7B □	1 FPS	1 FPS	54.20	8.60	63.50	42.10	47.70	70.70	57.60	46.60	74.50	71.30	61.40	33.20	40.80	43.10	39.03	47.51
FluxMem-7B □	1 FPS	1 FPS	–	–	–	–	59.60	70.70	63.00	53.40	**81.20**	**75.20**	67.20	–	–	–	–	53.30
Backbone Baseline and CIAct																		
Qwen3-VL-8B ○	1 FPS	1 FPS	55.60	12.90	63.50	44.00	59.60	73.30	58.70	57.30	69.80	71.30	65.00	**37.70**	40.40	60.80	46.30	51.77
+ CIAct (BLIP) △	1 FPS	64	59.60	16.67	**67.57**	**47.95**	61.47	**78.45**	69.02	**61.24**	73.83	72.28	69.38	24.07	60.42	56.60	47.03	54.79
+ CIAct (SigLIP2) △	1 FPS	64	**61.95**	12.90	66.22	47.02	**62.39**	76.72	69.57	60.11	78.52	74.26	**70.26**	25.93	61.25	63.28	50.15	**55.81**

Implementation Details. We use LLaVA-OneVision-7B (Li et al., 2024) and Qwen3-VL-8B (Yang et al., 2025a) as the base MLLMs of CIAct. The reference video frames used in Eq. 9 are sampled from LLaVA-Video (Zhang et al., 2024b) with a number of $16k$. Videos are processed at 1 FPS, and up to 64 observed frames are selected for answer generation. Across all experiments, we set the reference offset to $c = 0.10$ and the temporal decay factor to $\gamma = 0.95$. The response threshold and temporal factors are specified for each evaluation, and more details can refer to Appendix.

5.2 Quantitative Analysis

Comparison on the streaming benchmark. We first compare our CIAct to existing MLLMs and methods on the streaming benchmark of OVO in Tab 1. The first observation is that advanced streaming MLLMs achieve notable progress on all splits of streaming video understanding, mainly owing to the curation of fine-grained proactive response examples. For example, Streamo (Xia et al., 2026) achieves an overall score of 55.61% after being tuned by 465K proactive response examples. We can also observe that, while training-free methods generally present competitive results, their primary advantages lie in *real-time perception* and *backward tracing*, which are also evaluated under the semi-online setting as mentioned above. For example, QueryStream (Zhang et al., 2026a) achieves 61.40% on real-time perception but only 39.03% on forward active responding. Besides, these two splits exhibit a significant prior toward *recent frames* (Shen et al., 2026), reducing the challenges. In contrast, our CIAct achieves outstanding results across all splits using the most rigorous *proactive response* protocol. For example, CIAct with SigLIP2 achieves an overall score of 55.81%, slightly outperforming the SOTA Streamo at 55.61% without dedicated training. Considering its training-free setup, these results are in fact very remarkable.

Proactive Comparison on Offline Benchmarks. Table 2 evaluates whether proactive response can reduce video observation while preserving answer quality. It can be seen that the compared streaming models respond early during video watching, but their accuracy suffers from drastic declines compared the one using full video viewing. For instance, on LVBench, FluxMem[2] views only 1% of the video, but its accuracy drops by 16.4 points. This implies that their response time deviates greatly from the actual time. In contrast, although our CIAct consumes more viewing time, its performance does not differ much from the full viewing ones. Moreover, on the "needle-in-a-haystack" task, CIAct can achieve obvious gains while saving up to 38% viewing time. These significant results not only validate the proactive response capability of CIAct, but also confirm its merits in efficient and effective online video interaction.

Ablation Study. In the first block of Table 3, we ablate the effects and designs of CUSUM. Here, *temp. decay* denotes the temporal decay factor γ in Eq. 5. It can first be seen that direct thresholding

[2]The official code does not specify the threshold; we use 0.5.

Table 2: Comparison on offline video benchmarks under the strict proactive evaluation protocol. ○ and △ denote the *offline* and *proactive response* protocols, respectively.

Models	Frames	LongVideoBench		LVBench		Video-MME		MLVU	
		Score	Avg. viewed	Score	Avg. viewed	Score	Avg. viewed	Score	Avg. viewed
Dispider-7B ○	1 FPS	55.0	100%	41.3	100%	57.2	100%	61.7	100%
+ Proactive Evaluation △	1 FPS	48.5	30%	34.6	2%	49.7	19%	57.1	16%
TimeChat-Online-7B ○	1 FPS	57.7	100%	43.1	100%	62.5	100%	65.4	100%
+ Proactive Evaluation △	1 FPS	52.2	27%	35.3	8%	54.4	14%	56.7	6%
FluxMem-7B ○	1 FPS	61.1	100%	42.5	100%	65.3	100%	73.1	100%
+ Proactive Evaluation △	1 FPS	42.0	11%	26.1	1%	42.6	4%	46.9	5%
Backbone Baselines and CIAct									
LLaVA-OneVision-7B ○	1 FPS / 64	58.3	100%	41.0	100%	58.9	100%	67.7	100%
+ CIAct (SigLIP2) △	1 FPS / 64	58.3	79%	42.3	73%	57.5	81%	67.8	88%
+ CIAct (BLIP) △	1 FPS / 64	54.7	72%	41.1	62%	55.8	76%	66.5	88%
Qwen3-VL-8B ○	1 FPS / 64	63.1	100%	43.8	100%	67.6	100%	71.0	100%
+ CIAct (SigLIP2) △	1 FPS / 64	62.2	79%	44.2	73%	63.5	81%	70.8	88%
+ CIAct (BLIP) △	1 FPS / 64	59.9	71%	46.2	62%	63.8	73%	72.0	88%

Table 3: Ablation Study. Values in parentheses denote the average fraction of video viewed. All settings use 1 FPS sampling and up to 64 frames as input. ⋆ marks our final configuration.

Setting	LongVideoBench	MLVU
Accumulation Design		
Direct threshold	47.5 (15%)	55.8 (25%)
CUSUM	51.1 (37%)	56.1 (31%)
+ Cold start	54.6 (34%)	58.6 (28%)
+ Temp. decay	52.6 (43%)	58.0 (40%)
+ Both⋆	58.3 (79%)	67.8 (88%)
Response Threshold		
$\theta = 50$	57.3 (60%)	65.5 (72%)
$\theta = 70^{\star}$	58.3 (79%)	67.8 (88%)
$\theta = 90$	58.4 (92%)	68.2 (95%)

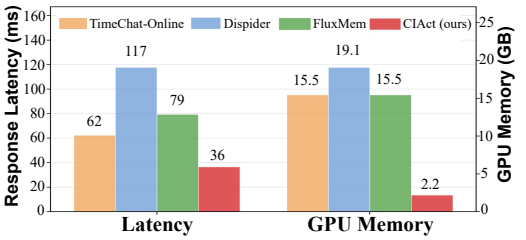


Figure 3: Comparison with existing methods in Response Latency and GPU memory overhead.

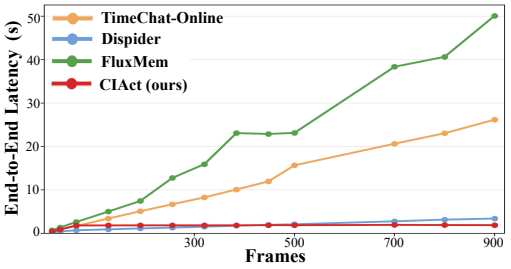


Figure 4: The latency of different methods with respect to the number of frames.

is hard to achieve proper proactive response capability for MLLMs, *e.g.*, achieving only 47.5 on LongVideoBench after viewing 15% of the video. It suggests that direct thresholding is merely a single-step estimation, which can not well reflect the cumulative states for proactive response. Thus, the use of the CUSUM function can greatly improve response accuracy and performance. Besides, it can be also seen that the cold-start design in CIAct is also very beneficial, particularly for long video understanding tasks, *e.g.*, +3.6% on LongVideoBench. And the temporal decay factor in Eq. 5 is also important for CUSUM. In addition to the obvious performance gains, this setting can prevent early and noise relevances from being indefinitely accumulated. And these results show that the performance gains of CIAct do not come from simply delaying the response, but instead from the progressive refinement of cumulative information until the response can be triggered.

The second block of Tab. 3 examines the sensitivity of the response threshold. It can be seen that this threshold is related to the viewing time of CIAct, but the performance gains of CIAct is not just from more viewing time. For instance, on LongVideoBench, increasing the threshold from $\theta = 70$ to $\theta = 90$ raises the average viewed fraction from 79% to 92%, but yields only a marginal gain of 0.1 points. This result suggests that the main merits of CIAct lies in its proper definition of proactive response as cumulative information estimation.

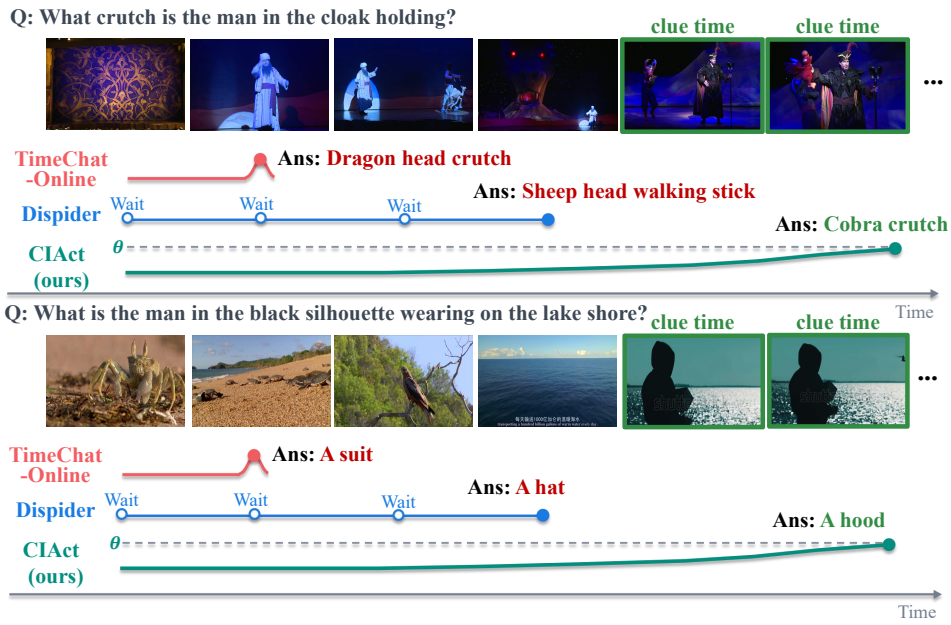


Figure 5: Visualized proactive responses of CIAct and the compared methods. Green boxes mark frames containing sufficient answer cues. In comparison, our CIAct merits in deciding the right time to response and the relevant frames for answering via its cumulative information estimation.

Efficiency Analysis. From Fig. 3, it can be seen that our CIAct has much lower response latency and peak GPU overhead than TimeChat-Online (Yao et al., 2025), Dispider (Qian et al., 2025) and FluxMem (Xie et al., 2026). As introduced above, the task definition in Eq. 6 of CIAct can enable it to use a lightweight VL model like SigLIP2 to independently measure query-frame information for response state estimation, rather than using all historical frames. Besides, the memory replay can also greatly reduce the input of video frames to MLLMs. In this case, from Fig. 4, we can also see that CIAct can help MLLMs to maintain a low but stable latency, without a significant increase as the video progresses. Overall, these results show that our CIAct can greatly reduce the proactive response costs of MLLMs while preserving high capabilities.

5.3 Qualitative Analysis

We further compare the proactive responses of TimeChat-Online, Dispider, and CIAct on the same video stream, as shown in Fig. 5. TimeChat-Online relies on visual feature changes between adjacent frames to trigger responses, and therefore tends to answer prematurely when an irrelevant scene transition occurs. Dispider uses a trained lightweight decision module to determine whether the current video clip supports an answer. However, it responds only after an explicit local cue appears and lacks the ability to accumulate weak evidence from earlier frames. In contrast, CIAct ignores the irrelevant scene transition and accumulates question-relevant evidence across frames with CUSUM, triggering a response only when the accumulated evidence becomes sufficient. These examples show that CIAct determines when to respond based on accumulated question-relevant information, rather than a single visual change or an instantaneous state prediction.

6 Conclusion

We presented CIAct, a training-free and plug-and-play framework that enables existing MLLM to respond proactively in online scenarios. CIAct reformulates response timing as cumulative information estimation, using lightweight VLMs to assess frame-query relevance, decayed CUSUM to accumulate evidence, and memory replay to retain relevant observations for efficient answer generation. This design requires no additional training and can be directly plugged into existing MLLMs. Extensive experiments on both streaming and offline benchmarks demonstrate that CIAct delivers strong proactive response performance while cutting video observation and response costs, and it consistently benefits existing offline MLLMs without sacrificing answer quality.

AI USE STATEMENT

We used generative AI tools only for organizing and restructuring existing code during the research process. Generative AI tools were not used for code generation, research ideation, methodology development, experimental design, data analysis, manuscript writing, or editing. All research ideas, technical contributions, experimental results, and the final manuscript were developed and prepared by the authors. We have reviewed all AI-assisted outputs and take full responsibility for the final content of this work.

REFERENCES

Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, Jiabo Ye, Xi Zhang, Tianbao Xie, Zesen Cheng, Hang Zhang, Zhibo Yang, Haiyang Xu, and Junyang Lin. Qwen2.5-vl technical report, 2025. URL https://arxiv.org/abs/2502.13923.

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 *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18407–18418. IEEE, 2024a.

Tao Chen, Kun Zhang, Qiong Wu, Xiao Chen, Chao Chang, Xiaoshuai Sun, Yiyi Zhou, and Rongrong Ji. Scaling the long video understanding of multimodal large language models via visual memory mechanism, 2026. URL https://arxiv.org/abs/2603.29252.

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, 2024b. URL https://arxiv.org/abs/2404.16821.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 24185–24198, 2024c.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yiming Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Kaipeng Zhang, Limin Wang, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling, 2025. URL https://arxiv.org/abs/2412.05271.

Xin Ding, Hao Wu, Yifan Yang, Shiqi Jiang, Qianxi Zhang, Donglin Bai, Zhibo Chen, and Ting Cao. Streammind: Unlocking full frame rate streaming video dialogue through event-gated cognition. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13448–13459. IEEE, 2025.

Yi Ding and Xue Li. Time weight collaborative filtering. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pp. 485–492, 2005.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

Junho Kim, Hosu Lee, James M. Rehg, Minsu Kim, and Yong Man Ro. Stride: When to speak meets sequence denoising for streaming video understanding. In *European Conference on Computer Vision*, pp. 38–56. Springer, 2026.

Junhyeok Kim, Min Soo Kim, Jiwan Chung, Jungbin Cho, Jisoo Kim, Sungwoong Kim, Gyeongbo Sim, and Youngjae Yu. Egospeak: Learning when to speak for egocentric conversational agents in the wild. *arXiv preprint arXiv:2502.14892*, 2025.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International conference on machine learning*, pp. 12888–12900. PMLR, 2022.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *Science China Information Sciences*, 68(10):200102, 2025.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, et al. Videochat3: Fully open video mllm for efficient and generalist video understanding. *arXiv preprint arXiv:2607.14935*, 2026.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. In *Proceedings of the 2024 conference on empirical methods in natural language processing*, pp. 5971–5984, 2024.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12585–12602, 2024.

Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18902–18913. IEEE, 2025.

Ewan S Page. Continuous inspection schemes. *Biometrika*, 41(1/2):100–115, 1954.

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. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24045–24055. IEEE, 2025.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Yujiao Shen, Shulin Tian, Jingkang Yang, and Ziwei Liu. A simple baseline for streaming video understanding. *arXiv preprint arXiv:2604.02317*, 2026.

Baiyang Song, Yuli Lin, Qiong Wu, Tao Chen, Jun Peng, Xiao Chen, Yiyi Zhou, and Rongrong Ji. Towards a dynamic and fixed-budget memory bank for efficient streaming video understanding, 2026. URL https://arxiv.org/abs/2606.25658.

Bo Tong, Bokai Lai, Yiyi Zhou, Gen Luo, Yunhang Shen, Ke Li, Xiaoshuai Sun, and Rongrong Ji. Flashsloth: Lightning multimodal large language models via embedded visual compression. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 14570–14581. IEEE, 2025.

Junlong Tong, Yao Zhang, Anhao Zhao, Yingqi Fan, Yunpu Ma, and Xiaoyu Shen. Proactivellm: Learning active interaction for streaming large language models. *arXiv preprint arXiv:2606.00523*, 2026.

Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025.

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. *Advances in Neural Information Processing Systems*, 38:132332–132359, 2026a.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, et al. Lvbench: An extreme long video understanding benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22958–22967. IEEE, 2025a.

Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Jiansheng Wei, Huishuai Zhang, and Dongyan Zhao. Videollm knows when to speak: Enhancing time-sensitive video comprehension with video-text duet interaction format, 2025b. URL https://arxiv.org/abs/2411.17991.

Yueqian Wang, Songxiang Liu, Disong Wang, Nuo Xu, Guanglu Wan, Huishuai Zhang, and Dongyan Zhao. Mmduet2: Enhancing proactive interaction of video mllms with multi-turn reinforcement learning. In *International Conference on Learning Representations*, volume 2026, pp. 23257–23270, 2026b.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2024.

Jiaer Xia, Peixian Chen, Mengdan Zhang, Xing Sun, and Kaiyang Zhou. Streaming video instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 31219–31229, 2026.

Yiweng Xie, Bo He, Junke Wang, Xiangyu Zheng, Ziyi Ye, and Zuxuan Wu. Fluxmem: Adaptive hierarchical memory for streaming video understanding. *arXiv preprint arXiv:2603.02096*, 2026.

Weicai Yan, Yuhong Dai, Qi Ran, Haodong Li, Wang Lin, Hao Liao, Xing Xie, Tao Jin, and Jianxun Lian. Proact-vl: A proactive videollm for real-time ai companions, 2026. *URL https://arxiv.org/abs/2603.03447*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025a.

Haolin Yang, Feilong Tang, Lingxiao Zhao, Xinlin Zhuang, Yifan Lu, Xiang An, Ming Hu, Xiaofeng Zhang, Abdalla Swikir, Junjun He, et al. Streamagent: Towards anticipatory agents for streaming video understanding. *arXiv preprint arXiv:2508.01875*, 2025b.

Linli Yao, Yicheng Li, Yuancheng Wei, Lei Li, Shuhuai Ren, Yuanxin Liu, Kun Ouyang, Lean Wang, Shicheng Li, Sida Li, et al. Timechat-online: 80% visual tokens are naturally redundant in streaming videos. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 10807–10816, 2025.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pp. 543–553, 2023.

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. *arXiv preprint arXiv:2406.08085*, 2024a.

Kairui Zhang, Zhenyu Yang, Bing Wang, Shengsheng Qian, and Changsheng Xu. Querystream: Advancing streaming video understanding with query-aware pruning and proactive response. In *The Fourteenth International Conference on Learning Representations*, 2026a.

Kecheng Zhang, Zongxin Yang, Mingfei Han, Haihong Hao, Yunzhi Zhuge, Changlin Li, Zhihui Li, Xiaojun Chang, et al. Progressive online video understanding with evidence-aligned timing and transparent decisions. In *International Conference on Learning Representations*, volume 2026, pp. 5092–5121, 2026b.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024b.

Yikai Zheng, Xin Ding, Yifan Yang, Shiqi Jiang, Hao Wu, Qianxi Zhang, Weijun Wang, Ting Cao, and Yunxin Liu. Em-garde: A propose-match framework for proactive streaming video understanding. *arXiv preprint arXiv:2603.19054*, 2026.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, et al. Mlvu: Benchmarking multi-task long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701. IEEE, 2025.

A Appendix

A.1 More Implementation Details.

A.1.1 Query–Frame Relevance

Given a question Q and a video frame V_t, CIAct computes their raw relevance as $s_t = s(Q, V_t)$. We instantiate the scoring function s with two interchangeable lightweight vision–language models. Both models remain frozen and require no additional training for proactive response.

Vision–language embedding. We use a pretrained SigLIP2 model (Tschannen et al., 2025) that maps the question and frame into a shared embedding space. Their relevance is measured by cosine similarity:

$$s_t = \frac{\phi_T(Q)^\top \phi_I(V_t)}{\|\phi_T(Q)\|_2 \, \|\phi_I(V_t)\|_2}, \tag{12}$$

where ϕ_T and ϕ_I denote the text and image encoders, respectively. A higher score indicates stronger semantic alignment between the question and frame.

Image–text matching. We alternatively use a pretrained BLIP model (Li et al., 2022) with an image–text matching (ITM) head. BLIP jointly processes the question and frame through cross-modal interaction and predicts matched and mismatched logits. We define their difference as the relevance score:

$$s_t = \ell_{\text{match}}(Q, V_t) - \ell_{\text{mismatch}}(Q, V_t), \tag{13}$$

where ℓ_{match} and ℓ_{mismatch} denote the two output logits. A higher score indicates that BLIP is more confident that the frame matches the question.

For either implementation, we apply the same scoring function to the streaming frames and the external reference bank. The resulting raw scores therefore share the background calibration, evidence accumulation, and memory replay procedures described in the main text.

A.1.2 Experimental Settings

Across all experiments, videos are processed at 1 FPS, and CIAct uses a reference offset of $c = 0.10$, a temporal decay factor of $\gamma = 0.95$, and a maximum memory budget of $K = 64$ frames. The remaining parameters are summarized in Table 4.

Table 4: Parameter configurations of CIAct. A single configuration is shared across all benchmarks within each evaluation setting.

Setting	Relevance Model	θ	h	$\omega_{\min}$
Streaming benchmark	SigLIP2 / BLIP	20	40	0.30
Offline benchmarks	SigLIP2	70	80	0.50
Offline benchmarks	BLIP	350	80	0.50

The response threshold θ is configured separately for different relevance models because background standardization aligns their marginal score distributions but does not eliminate differences in temporal evidence dynamics. We therefore use different thresholds for SigLIP2 and BLIP while keeping each configuration fixed across all four offline benchmarks.

The two evaluation settings also require different temporal configurations. In the streaming setting, questions may arrive at arbitrary timestamps and historical evidence is replayed at query time; hence, a lower threshold supports timely responses. We also use a shorter half-life and a lower weight floor to prioritize frames close to the response time. For offline benchmarks, questions are provided at the beginning and relevant evidence may be distributed over a longer temporal range. Accordingly, a longer half-life and a higher weight floor preserve broader historical evidence for answer generation.

A.2 ALGORITHM

Algorithm 1 summarizes the complete online inference procedure of CIAct. A question arriving at t_q first triggers background calibration and historical evidence replay. If the historical evidence is insufficient, CIAct continues updating the cumulative state over incoming frames until a response is triggered.

Algorithm 1: Online inference with CIAct

Input: Video stream $V = \{V_t\}_{t=1}^{T}$; question Q arriving at t_q; relevance function s; reference bank $\mathcal{V}_{\text{ref}}$ with M frames; answering MLLM $\mathcal{G}$.
Parameters: Reference offset c, temporal decay γ, response threshold θ, temporal half-life h, minimum temporal weight $\omega_{\min}$, and memory budget K.
Output: Response time t_r and answer $\hat{y}$.
Compute μ_Q and σ_Q on $\mathcal{V}_{\text{ref}}$ using Eq. 9.
Set $E \leftarrow 0$ and $historical_hit \leftarrow$ false.
// Replay the history available when the question arrives.
for $t = 1, \ldots, t_q$ **do**
    $s_t \leftarrow s(Q, V_t)$ and $z_t \leftarrow (s_t - \mu_Q)/\sigma_Q$.
    $E \leftarrow \max(0, \gamma E + z_t - c)$.
    if $E \geq \theta$ **then** $historical_hit \leftarrow$ true.
end for
if $historical_hit$ **then** $t_r \leftarrow t_q$.
else continue observing future frames.
    for $t = t_q + 1, \ldots, T$ **do**
        $s_t \leftarrow s(Q, V_t)$ and $z_t \leftarrow (s_t - \mu_Q)/\sigma_Q$.
        $E \leftarrow \max(0, \gamma E + z_t - c)$.
        if $E \geq \theta$ **then** $t_r \leftarrow t$ and **break**.
    end for
    if no threshold crossing occurs **then** $t_r \leftarrow T$.
end if
// Recall a compact visual memory for answering.
for $t = 1, \ldots, t_r$ **do**
    $w_t \leftarrow \max(z_t, 0)[\omega_{\min} + (1 - \omega_{\min})2^{-(t_r - t)/h}]$.
end for
$V_m \leftarrow \text{TopK}_{V_t \in V_{1:t_r}}(w_t)$; arrange the selected frames in temporal order.
$\hat{y} \leftarrow \mathcal{G}(V_m, Q)$ and **return** $(t_r, \hat{y})$.