ASVBENCH: ASYNCHRONOUS RESPONSE TO FINE-GRAINED TRIGGERS IN STREAMING VIDEOS

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming video understanding is a pivotal technology for bridging Multimodal Large Language Models (MLLMs) with real-world AI hardware, such as AR glasses and embodied robots. However, current benchmarks are largely confined to a simplistic "immediate-query, immediate-answer" scenario. We argue that real-world interactions demand a more advanced capability for **Asynchronous Response**, where a model's output is delayed until specific visual conditions are met within a maximum temporal window. This paradigm fundamentally necessitates **fine-grained understanding**, as the model must vigilantly track the subtle, object-centric details that act as the visual triggers. This critical interplay between asynchronous response and fine-grained triggers has been largely overlooked. Therefore, we propose **ASVBench**, comprising 1452 videos and 4350 challenging Q&A pairs, as the first benchmark for evaluating Asynchronous Response to Fine-grained Triggers in streaming video. Specifically, we build a dual-axis evaluation framework for achieving this goal: (1) The temporal axis assesses models across three core scenarios: Immediate Q&A (responding to the present visual state), Prospective Q&A (responding when future visual triggers are met), and Continuous Q&A (continuously monitoring if triggers are met). (2) The cognitive axis requires the model to respond to visual triggers at different granularities, including recognition-level, comprehension-level, and reasoning-level triggers. ASVBench evaluation results reveal a significant gap between existing MLLMs and human performance, with accuracy plummeting as asynchronous temporal complexity increases. Detailed diagnostics indicate that current models suffer from micro-feature over-smoothing, memory eviction of latent goals, and impatience-driven premature responses, with early errors cascading in multi-shot streams. These insights expose critical bottlenecks in asynchronous responses to fine-grained triggers, providing a roadmap for next-generation streaming MLLMs.

1 INTRODUCTION

Streaming video understanding (Zhao et al., 2023; Qian et al., 2024; Foo et al., 2023) is crucial to bridge Multimodal Large Language Models (MLLMs) (Li et al., 2023; Achiam et al., 2023; Zhang et al., 2023) with real-world hardware such as AR glasses and embodied robots. Because the reasoning capabilities of these devices are strictly bottlenecked by continuous environmental perception, the fidelity of streaming analysis directly dictates overall system reliability. Consequently, robust streaming understanding is a fundamental prerequisite for deploying safe and effective AI applications in open, uncontrolled environments.

However, current benchmarks (Chen et al., 2024; Huang et al., 2025; Lin et al., 2026; Song et al., 2024; Yuan et al., 2025; Zhang et al., 2025) for streaming video understanding are confined to a simplistic "immediate-query, immediate-answer" scenario (Fig. 1 top). This paradigm, where a model must react instantly to the current visual input, fails to address the complexities of real-world interactions. In practice, streaming scenarios demand an Asynchronous Response capability, where a model must generate an output only when specific visual fine-grained triggers are met after a query has been posed within a maximum temporal window (Prospective Q&A and Continuous Q&A of Fig. 1 bottom). For instance, in Continuous Q&A part of Fig. 1, a model might be asked to identify the purpose of applying alcohol (at 2:40), but the visual evidence required to answer (i.e., alcohol and cotton swabs) only appears nearly thirty seconds later (at 3:05 and 3:26). Similarly, a

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.

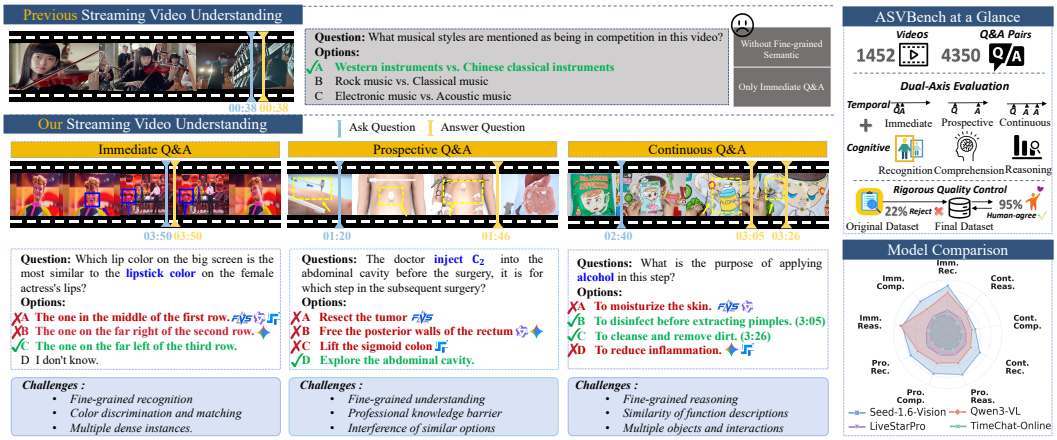


Figure 1: **Comparison of ASVBench with previous streaming video understanding benchmarks.** Left: prior benchmarks are confined to "immediate-query, immediate-answer" scenarios, while ASVBench evaluates asynchronous response to fine-grained triggers across Immediate, Prospective, and Continuous Q&A. Right: the "ASVBench at a Glance" panel summarizes dataset statistics and a radar chart compares models across the 9 sub-metrics, revealing a large offline-streaming gap on both accuracy and timing.

smart oven tasked with "notify me when the cake is ready", should only send a notification when it visually determines the cake has turned golden brown, rather than continuously reporting the baking status. Thus, the ability to retain a user instruction as a latent goal and trigger a response only when future fine-grained cues fulfill the conditions for answering, or safely timeout if they do not, is crucial for real-world applications, yet remains largely unexplored in streaming video understanding benchmarks for this kind of interaction.

To address this critical evaluation gap, we propose **ASVBench**, the first benchmark designed to assess Asynchronous Response to Fine-Grained Triggers in streaming video, as shown in Table 1. The dual-axis design of ASVBench is crafted to tackle the challenges outlined above. The temporal axis directly evaluates models across different asynchronous scenarios, capturing their macro-level performance in three modes: Immediate Q&A for immediately responding to the current visual state, Prospective Q&A for holding the latent goal to respond when future visual triggers are met, and Continuous Q&A for continuously holding the goal and responding based on ongoing monitoring. Meanwhile, the cognitive axis systematically varies the granularity of the triggers, including recognition-level, comprehension-level, and reasoning-level, to assess model adaptation to increasing complexity of the visual trigger.

Table 1: Comparison between ASVBench and existing video benchmarks. Videos/QA: number of videos / Q&A pairs. St.: streaming. FG: fine-grained triggers. Asy.: asynchronous response. Co.: continuous monitoring. TS: timestamp-aware queries. ET: event-timing evaluation.

Benchmark	Videos/QA	St.	FG	Asy.	Co.	TS	ET
Offline Video Understanding							
Video-MME (Fu et al., 2025)	900/2.7K	×	×	×	×	×	×
TemporalBench (Cai et al., 2024)	2K/10K	×	✓	×	×	×	×
VideoRefer (Yuan et al., 2025)	198/1K	×	✓	×	×	×	×
Streaming Video Understanding							
StreamingBench (Lin et al., 2026)	900/4.5K	✓	×	×	✓	✓	×
OVBench (Huang et al., 2025)	1.5K/7K	✓	×	×	✓	✓	×
OVO-Bench (Li et al., 2025)	644/2.8K	✓	×	✓	×	✓	×
SVBench (Yang et al., 2025b)	1.4K/50K	✓	×	×	✓	✓	×
TV-Online (Qian et al., 2026)	50K/200K	✓	×	✓	✓	✓	×
OmniPro (Zhao et al., 2026)	1.3K/2.7K	✓	×	✓	✓	✓	✓
ProactiveBench (Du et al., 2026)	911/911	✓	×	✓	✓	✓	✓
ASVBench (Ours)	1.5K/4.4K	✓	✓	✓	✓	✓	✓

We conduct a comprehensive evaluation of 20 state-of-the-art MLLMs on ASVBench. To estimate an upper bound for streaming video understanding, we evaluate powerful offline general-purpose models, including GPT-5.5 (OpenAI, 2026), Claude Opus 4.6 (Anthropic, 2026), Kimi-K2.5 (Kimi Team, 2026), and GLM-5 (GLM-5 Team, 2026), under a causal sliding window protocol that preserves temporal causality. These are systematically compared against native streaming-specialized architectures such as MiniCPM-o-4.5 (Cui et al., 2026), LiveStarPro (Yang et al., 2026), and Flash-VStream (Zhang et al., 2025). Our results reveal a significant gap compared to human performance, with accuracy plummeting as asynchronous temporal complexity increases. Most notably, native

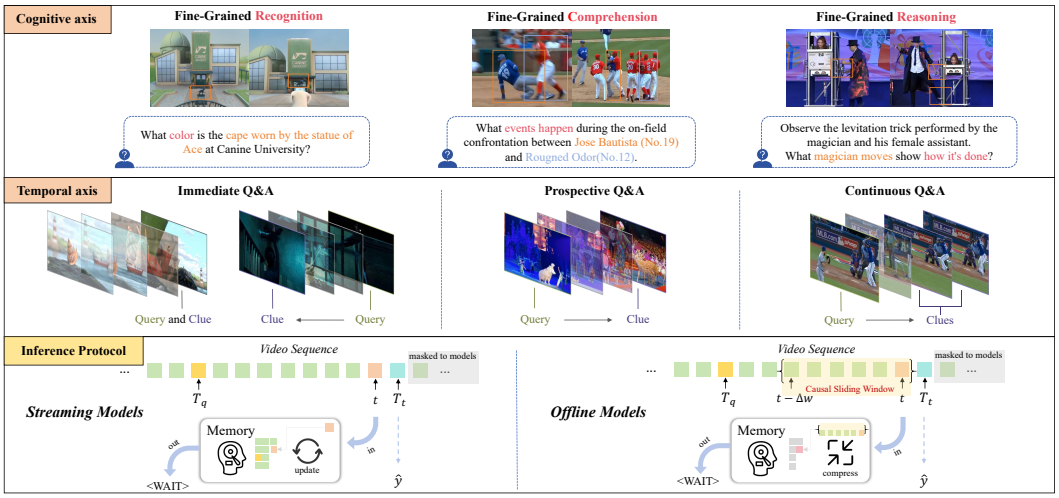


Figure 2: **The dual-axis evaluation framework and the inference protocol of ASVBench.** The cognitive axis spans fine-grained recognition, comprehension, and reasoning, while the temporal axis defines Immediate, Prospective, and Continuous Q&A based on query-clue timing. The bottom contrasts inference protocols: continuous processing for streaming models versus causal sliding windows for offline models that never see future frames.

streaming models exhibit a severe performance collapse and consistently underperform the adapted offline baselines by a wide margin on ASVBench.

We further attribute this performance collapse to four fundamental architectural bottlenecks. The spatial bottleneck involves an aggressive over-smoothing of micro-features during real-time compression. The temporal bottleneck is driven by severe latent goal decay in asynchronous scenarios, where memory constraints progressively discard unfulfilled instructions over time, causing models like GPT-5.5 and GLM-5 to plummet to 58.81 and 52.09 on Continuous Q&A. The behavioral bottleneck reveals a prevalent impatience bias, which triggers premature hallucinations well before actual visual cues appear. Finally, models suffer from cascading failures in multi-shot streams, where flawed early states continuously propagate and impair ongoing context tracking.

In summary, our contributions are as follows.

- We introduce ASVBench, the first benchmark dedicated to evaluating asynchronous responses to fine-grained visual triggers. To systematically assess this capability, we design a comprehensive dual-axis framework spanning temporal and cognitive dimensions.
- We propose a causal sliding window protocol and an evaluation framework combining the Semantic Score and the Joint Asynchronous Score. Timing credit is counted only when the answer is semantically correct.
- By evaluating 20 state-of-the-art MLLMs, we identify four fundamental streaming bottlenecks: spatial over-smoothing, temporal goal decay, impatience bias, and cascading multi-shot errors. These diagnostic insights establish a clear roadmap for future designs.

2 Related Work

Offline Video Understanding. Recent MLLMs (Comanici et al., 2025; OpenAI, 2026; Anthropic, 2026; Bai et al., 2025; ByteDance, 2025; Yu et al., 2025) have substantially advanced offline video comprehension by encoding full sequences into visual tokens and modeling complex spatiotemporal relations. Complementary benchmarks further stress long-form understanding (Fu et al., 2025), fine-grained temporal reasoning (Cai et al., 2024), and object-centric perception (Yuan et al., 2025). However, these settings assume complete access to the entire video at inference time, which fundamentally limits their applicability to real-time streaming.

Streaming Video Understanding. Streaming video understanding requires online reasoning from past and present frames. Recent architectures enhance real-time efficiency via memory compres-

sion, token dropping, and proactive decoding (Zhang et al., 2025; Yao et al., 2025; Xiong et al., 2025; Huang et al., 2025; Cui et al., 2026; Yang et al., 2026; Li et al., 2026), occasionally integrating streaming speech (Chen et al., 2025) or operating entirely memory-free (Shen et al., 2026). Accordingly, existing benchmarks predominantly evaluate continuous immediate Q&A or multi-turn dialogue (Lin et al., 2026; Huang et al., 2025; Li et al., 2025; Yang et al., 2025b). While recent efforts advance toward query-aware pruning (Zhang et al., 2026; Qian et al., 2026) and proactive temporal responding (Zhao et al., 2026; Du et al., 2026), they remain confined to macroscopic events, as shown in Table 1. Crucially, they fail to evaluate the retention of long-horizon latent goals bound to micro-feature visual triggers. ASVBench addresses this fundamental gap by explicitly assessing asynchronous scheduling conditioned on fine-grained visual cues, introducing a Joint Asynchronous Score that rigorously couples semantic accuracy with precise response timing.

3 THE FRAMEWORK OF ASVBENCH

ASVBench adopts a dual-axis framework evaluating fine-grained perception and asynchronous response. We formalize the task across temporal and cognitive dimensions, establishing inference protocols for both native streaming and offline models alongside quantitative metrics.

3.1 EVALUATION DIMENSIONS

To evaluate asynchronous response capability in streaming video, we categorize tasks along the temporal dimension based on query-clue timing. As illustrated in Figure 2, questions are classified into Immediate Q&A, Prospective Q&A, and Continuous Q&A. Within Immediate Q&A, we differentiate present versus past states based on clue positioning. Furthermore, to evaluate fine-grained visual triggers, our framework spans three cognitive dimensions: recognition, comprehension, and reasoning about one and the same visual event.

Dimension 1: Fine-Grained Recognition. This dimension assesses the model's ability to identify specific objects, actions, or attributes. The objective is to extract explicit visual facts that are definitively resolved the moment the corresponding visual trigger appears.

Dimension 2: Fine-Grained Comprehension. This dimension assesses the tracking and interpretation of event progression, requiring models to capture key moments and interactive processes. It demands a holistic understanding of how dynamics and spatial-temporal interactions unfold.

Dimension 3: Fine-Grained Reasoning. This dimension focuses on inferring causal links and temporal relationships across sequences of actions. Since the required answer is rarely depicted explicitly in any single frame, the model must synthesize disparate visual clues across the temporal stream to deduce underlying causes, intentions, or logical outcomes.

3.2 TASK FORMULATION

Unlike offline settings where the entire video sequence is accessible at once, streaming evaluation rigorously enforces strict causality. We define the basic task as processing a sequence of video frames $V \in \mathbb{R}^{N\times W\times H\times C}$. In our streaming protocol, when a query x is posed at a specific timestamp T_q, the system feeds subsequent video frames forward with a fixed step size of $\Delta t = 1$ second. Per-frame stepping adds latency and redundant frames; a one-second step matches the granularity of real-time interaction. The model must dynamically process the incoming stream until it generates the final answer $\hat{y}$. Based on this formulation, we establish rigorous evaluation pipelines tailored to different model paradigms, both streaming and offline.

3.2.1 INFERENCE PROTOCOL FOR STREAMING MODELS

Streaming models process video on the fly while maintaining a persistent internal memory state. Let M_t denote the model's memory state at timestamp t. When a query x is posed at T_q, it is encoded into the initial memory M_{T_q}. For any subsequent timestamp $t \geq T_q$, the model receives the incoming video frame v_t. The streaming inference involves continuous state updates and trigger-driven response generation. Specifically, the model updates its memory cache recurrently via $M_t = \text{Update}(M_{t-\Delta t}, v_t)$. Based on M_t, the model decides whether the fine-grained visual trigger has

appeared. It must either output a <WAIT> signal to indicate the absence of the trigger and continue monitoring, or generate the final response $\hat{y}$ upon trigger detection.

3.2.2 Inference Protocol for Offline Models

While streaming models maintain internal states natively, evaluating offline MLLMs requires simulated streaming to rigorously enforce causality. Therefore, we define a causal sliding window Δw. The available visual context at any inference step t is formulated as:

$$\mathcal{V}_t = V([t - \Delta w, t]) \tag{1}$$

where $\mathcal{V}_t$ represents the most recent visual segment up to the current timestamp t. Future frames where $\tau > t$ are strictly masked to prevent causal leakage.

At any inference step t, the system maintains a cumulative memory state M_t to explicitly preserve historical cache. Specifically, the dynamic visual segment $\mathcal{V}_t$ within the current sliding window is compressed into a condensed representation $c_t = \text{Compress}(\mathcal{V}_t)$. The model's memory is recurrently updated via $M_t = M_{t-\Delta t} \cup \{c_t\}$. The model then evaluates the static query x against this updated M_t at every interval Δt, outputting a <WAIT> signal to maintain the latent goal, or the final answer $\hat{y}$ upon trigger detection.

To measure temporal tracking complexity, we categorize the evaluation into three asynchronous modes of Immediate, Prospective, and Continuous response:

- Mode 1: Immediate Q&A. The query x is posed at T_q, and the model must answer immediately based on the current or past visual state. The causal context relies on the current window $\mathcal{V}_t = V([T_q - \Delta w, T_q])$ for present clues, or ends at a past event timestamp $T_e \leq T_q$ where $\mathcal{V}_t = V([T_e, T_q])$.
- Mode 2: Prospective Q&A. The query x is posed at T_q, but the visual trigger occurs at a future timestamp T_t ($T_q < T_t$). Evaluated sequentially at $t \in \{T_q, T_q + \Delta t, \dots\}$, the model must output a semantic <WAIT> state for all $t < T_t$, and generate the final answer $\hat{y}$ only when $t \geq T_t$. Each clip is sliced from the earliest query to the latest event (Step 6), defining the maximum horizon $T_{\max} = T_{\text{end}} - T_q$. Unanswered triggers by T_{max} require a rejection output, scored as a timeout.
- Mode 3: Continuous Q&A. We use two sequential visual triggers at T_{t1} and T_{t2} ($T_q < T_{t1} < T_{t2}$), indexed by $k \in \{1, 2\}$. At any step t, the model processes $\mathcal{V}_t$. It must output <WAIT> initially ($t < T_{t1}$) and during the inter-trigger monitoring phase ($T_{t1} < t < T_{t2}$). Upon detecting the triggers at $t = T_{t1}$ and $t = T_{t2}$, it should generate the intermediate answer $\hat{y}_1$ and final answer $\hat{y}_2$, respectively. An unanswered trigger at $T_{\max}$ is a timeout.

3.2.3 Evaluation Objective

We report a Semantic Score and a Joint Asynchronous Score. The **Semantic Score** S measures answer correctness via exact match:

$$S = \frac{1}{N} \sum_{i=1}^{N} \mathbb{I}(y_i = \hat{y}_i) \times 100 \tag{2}$$

where N is the number of instances and $\mathbb{I}(\cdot)$ yields 1 if the predicted answer $\hat{y}_i$ matches the ground truth y_i. For Continuous Q&A, a trial is correct only if both the intermediate and final predictions match ($\hat{y}_1 = y_1$ and $\hat{y}_2 = y_2$).

The **Joint Asynchronous Score** $\mathcal{J}$ represents the joint credit for a response that is both semantically correct and timely. Timing credit is counted only when the semantic prediction is correct, so a timely hallucination receives zero. For instance i and trigger k, let $T_{p,i,k}$ be the first time the model emits a non-<WAIT> answer for that trigger and $T_{t,i,k}$ the ground-truth trigger time, both in seconds:

$$\Delta T_{i,k} = T_{p,i,k} - T_{t,i,k}. \tag{3}$$

With decay coefficient $\alpha = 0.2\,\text{s}^{-1}$, the per-trigger coefficient is

$$\omega_{i,k} = \begin{cases} 0 & \Delta T_{i,k} < 0 \text{ (premature)} \\ e^{-\alpha \Delta T_{i,k}} & \Delta T_{i,k} \geq 0 \text{ (on time or delayed)} \\ 0 & \text{unanswered by } T_{\max}. \end{cases} \tag{4}$$

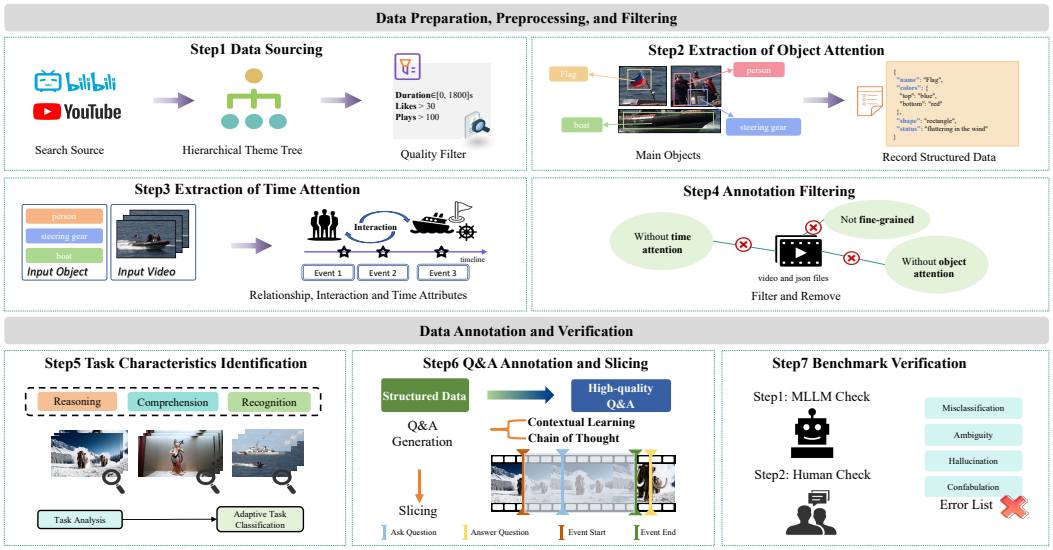


Figure 3: **The ASVBench-Miner pipeline supporting ASVBench.** This two-phase pipeline first filters raw videos and extracts object and temporal annotations (Steps 1-4). Phase 2 then leverages these annotations to identify tasks, generate timestamped Q&A pairs, slice videos, and rigorously verify the final dataset (Steps 5-7).

Prospective Q&A has one trigger, so $\omega_i = \omega_{i,1}$. Continuous Q&A has two triggers, and ω_i is their average, $(\omega_{i,1} + \omega_{i,2})/2$. The joint score is

$$\mathcal{J} = \frac{1}{N} \sum_{i=1}^{N} \left(\mathbb{I}(y_i = \hat{y}_i) \times \omega_i \right) \times 100. \tag{5}$$

Premature responses are zeroed, and delayed responses decay exponentially. A semantically wrong answer receives no credit.

4 The Data Engine of ASVBench

We develop ASVBench-Miner, an automated pipeline for video mining and annotation, as shown in Figure 3, from mining through verification.

4.1 Preparation, Preprocessing, and Filtering

Step 1: Data Sourcing. We source candidate videos from YouTube and BiliBili using a domain-oriented search guided by a hierarchical theme tree that narrows from broad domains to specific themes and keywords. A quality filter based on engagement and duration retains only videos with at least 30 likes and 100 plays, capped at 1800 seconds.

Step 2: Extraction of Object Attention. GPT-4o (Hurst et al., 2024) identifies and tracks fine-grained central objects under intensive interaction, recording their bounding timestamps and key physical properties, such as name, color, shape and state, into a structured format, which provides visual anchors for downstream interaction tracking.

Step 3: Extraction of Time Attention. Given the raw video and extracted object data, GPT-4o identifies fine-grained events, object interactions, and their exact appearance and disappearance timestamps, mapping contextual relationships to provide temporal attention for downstream reasoning.

Step 4: Annotation Filtering. We use Qwen-VL-Max (Bai et al., 2023) to filter the video together with its object- and time-attention annotations, removing videos that are not fine-grained or lack either attention type. This ensures only high-quality videos with both object and time attention remain for the later annotation stages.

Model	Immediate Q&A			Prospective Q&A			Continuous Q&A			$\mathcal{J}$	Overall
	REC.	COM.	REA.	REC.	COM.	REA.	REC.	COM.	REA.		
Human	**96.15**	**95.73**	**95.78**	**96.37**	**94.10**	**95.47**	**95.46**	**95.27**	**94.63**	–	**95.47**
GPT-4o (Text-Only)	29.43	27.58	33.42	29.15	31.46	32.13	32.74	30.88	29.70	–	30.64
Proprietary General-Purpose MLLMs											
Gemini-2.5-Pro Comanici et al. (2025)	90.76	72.68	82.40	67.35	66.12	66.67	48.16	42.45	48.22	–	68.97
Gemini-2.5-Flash Comanici et al. (2025)	75.71	62.54	82.62	58.42	59.56	58.46	35.51	30.58	44.67	–	64.79
Step-1o-Turbo-V StepFun (2025)	65.55	62.11	76.89	–	–	–	–	–	–	–	–
Seed-1.6-vision ByteDance (2025)	89.06	78.25	85.41	78.62	77.43	79.17	68.04	51.38	53.71	–	68.64
Claude Opus 4.6 Anthropic (2026)	90.42	78.36	**87.54**	79.23	78.17	80.60	68.93	52.21	**54.47**	–	69.83
GPT-5.5 OpenAI (2026)	**91.57**	**79.41**	86.30	**80.46**	**78.59**	**81.24**	**69.37**	**53.11**	53.96	–	**71.31**
Open-Source General-Purpose MLLMs											
MiniCPM-o-2.6 (8B) Team (2025a)	76.58	70.14	85.40	58.42	53.55	64.62	46.12	43.17	49.24	–	67.66
MiniCPM-V-4.5 (9B) Yu et al. (2025)	76.70	70.85	**88.72**	57.39	56.83	58.97	42.86	44.60	48.73	–	67.95
Keye-VL-1.5 (8B) Yang et al. (2025a)	76.33	68.59	86.14	30.24	38.52	44.10	23.27	20.14	21.32	–	65.35
Qwen3-VL (8B) Team (2025b)	76.21	59.72	88.17	63.23	56.83	56.41	48.98	44.96	51.27	–	65.85
Qwen3-VL (30B-A3B) Bai et al. (2025)	83.12	64.37	84.61	66.42	60.53	59.81	50.26	46.30	52.87	–	67.43
Kimi-K2.5 (1T-A32B) Kimi Team (2026)	83.44	70.61	82.57	64.51	62.28	63.17	49.42	47.29	49.26	–	62.45
GLM-5 (744B-A40B) GLM-5 Team (2026)	**84.23**	**72.59**	85.48	**68.53**	**64.40**	**65.13**	**51.29**	**50.38**	**54.61**	–	**68.31**
Open-Source Streaming-Specialized MLLMs											
Flash-VStream (7B) Zhang et al. (2025)	22.18	32.82	20.73	18.67	32.57	32.05	29.80	34.17	**46.19**	36.51	38.75
VideoChat-Online (4B) Huang et al. (2025)	26.53	35.12	22.38	28.46	31.22	29.51	25.87	26.10	24.39	34.20	27.68
StreamChat (7B) Xiong et al. (2025)	43.15	41.86	40.24	42.53	39.60	38.49	**39.81**	**37.13**	36.52	40.91	39.94
TimeChat-Online (7B) Yao et al. (2025)	36.29	38.71	33.54	32.40	28.65	30.07	27.53	25.96	26.47	23.14	30.64
MiniCPM-o-4.5 (9B) Cui et al. (2026)	**45.32**	**43.16**	**41.52**	43.29	40.54	38.67	37.80	35.29	34.92	41.59	**40.01**
LiveStarPro (8B) Yang et al. (2026)	44.21	42.57	40.13	**44.36**	**41.28**	**40.91**	38.64	34.63	36.89	**43.27**	38.58
VideoChat3 (4B) Li et al. (2026)	40.37	39.52	38.74	39.06	38.72	30.59	27.20	24.13	25.17	36.88	31.47

Table 2: **Main Results**. The best results are **bolded**. "–" means that the model does not support the certain input form. $\mathcal{J}$ denotes the Joint Asynchronous Score, which counts timing only when the answer is semantically correct.

4.2 Annotation and Verification

Step 5: Task Characteristics Identification. Based on the fine-grained object and event annotations, we use Claude-3.5-Sonnet (Anthropic, 2024) to analyze each video's content to assign a suitable task category: Reasoning, Comprehension, or Recognition.

Step 6: Q&A Annotation and Slicing. Prompted by Gemini-2.5-Pro (Comanici et al., 2025) using object and time attention, we generate Q&A pairs via Chain of Thought (Wei et al., 2022). For each instance, we mark four critical timestamps: Ask Question, Answer Question, Event Start, and Event End, and slice the video between the earliest and latest markers to remove redundant frames. The Answer Question timestamp defines the ground-truth trigger T_t for temporal-semantic alignment.

Step 7: Benchmark Verification. Professional annotators and Qwen3-Omni (Xu et al., 2025) perform multi-stage verification, filtering 5,600 Q&A pairs to 4,350 (22.32% rejection). Three annotators independently re-label 700 sampled pairs, yielding Fleiss' Kappa of 0.91, and our Human Baseline reaches 95.47% agreement (Table 2), jointly confirming objectivity and absence of AI hallucinations. The 1,250 rejected pairs fall into Visual Ambiguity (57%), Subjective Interpretation (21%), and LLM Hallucination (22%). Full protocols are in Appendix B.

5 Experiments

5.1 Experimental Setup

Dataset Statistics. ASVBench comprises 4,350 Q&A pairs from 1,452 videos across diverse domains. To accurately reflect real-world complexity, the videos span highly diverse, high-frequency physical domains, including egocentric views, artistic performances, medical scenarios, and daily life, as detailed in Appendix C.

Evaluated Models. We evaluate 20 leading proprietary and open-source general-purpose MLLMs, along with the latest streaming-specialized MLLMs, comparing them against a Human Baseline. Specifically, our evaluation includes: (1) Widely-used proprietary general-purpose models, for example GPT-5.5 (OpenAI, 2026), Claude Opus 4.6 (Anthropic, 2026), and Step-1o-Turbo-V (StepFun, 2025); (2) Prominent open-source general-purpose models, including GLM-5 (GLM-5

Team, 2026), Kimi-K2.5 (Kimi Team, 2026), and MiniCPM-V-4.5 (Yu et al., 2025); (3) The latest streaming-specialized models such as MiniCPM-o-4.5 (Cui et al., 2026), LiveStarPro (Yang et al., 2026), and TimeChat-Online (Yao et al., 2025).

Offline general-purpose models serve as a perception upper bound, adapting them via the causal sliding window protocol separates perception failures from streaming-specific flaws. We set $\Delta w = 32$ frames for all offline models; the ablation study (Appendix D) shows that $\Delta w = 64$ yields only marginal accuracy gains at the cost of prohibitive KV-cache memory and latency, breaking real-time constraints of live online interaction.

5.2 Main Results

We present the main results in Table 2, covering all models across three task modes. The text-only GPT-4o baseline reaches 30.64%, consistent with the chance level of the four-option and true/false questions, confirming that ASVBench eliminates language bias. Details are given in Appendix C.1.

Our experiments yield four key findings:

Finding 1: A Significant Gap to Human Performance. The Human Baseline achieves an overall accuracy of 95.47%, whereas the top model GPT-5.5 reaches an accuracy of only 71.31%, confirming that fine-grained streaming understanding remains an open challenge.

Finding 2: Performance Collapse in Asynchronous Scenarios. Accuracy degrades with temporal complexity. Models perform best in Immediate Q&A, drop in Prospective Q&A, and collapse in Continuous Q&A, where GPT-5.5 and GLM-5 plummet to 58.81 and 52.09, revealing sustained temporal tracking as the dominant bottleneck.

Finding 3: Streaming-Specialized Models Lag in Asynchronous Tracking. Streaming-specialized models such as LiveStarPro consistently underperform offline general-purpose models adapted via sliding windows, outputting collapsed answers over long horizons. Root causes are diagnosed in Section 5.3 of the experiments.

Finding 4: Impatience Bias in Response Timing. $\mathcal{J}$ exposes a severe Impatience Bias in streaming-specialized MLLMs. Premature responses with $\Delta T_{i,k} < 0$ receive no credit, and a timely answer counts only when it is semantically correct. LiveStarPro scores 43.27 and VideoChat3 scores 36.88. Future architectures may require Asynchronous RLHF to align patience with real-world temporal dynamics of an ongoing video stream.

5.3 Diagnostic Analysis of Streaming Failures

To diagnose the architectural bottlenecks behind streaming failures on ASVBench, we conduct three targeted analyses of where streaming fails.

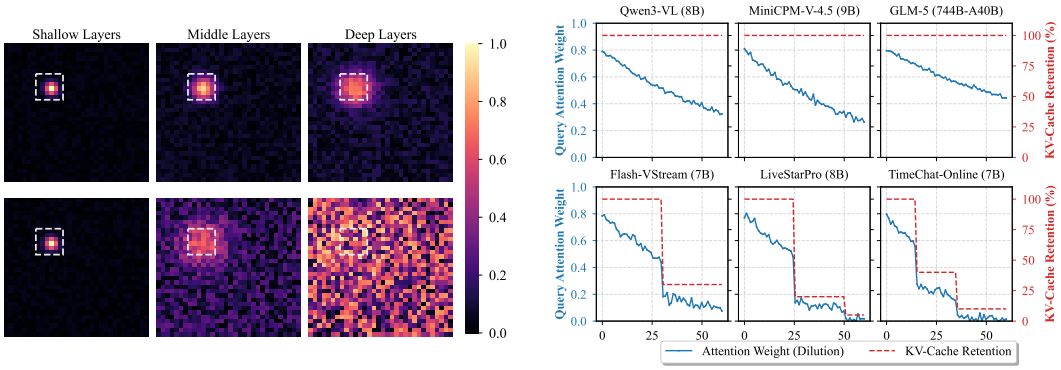
5.3.1 Spatial Failure: Over-Smoothing of Micro-Triggers

We analyze averaged attention maps from three general-purpose models, Qwen3-VL (8B), MiniCPM-V-4.5, and Keye-VL-1.5 (8B), and three streaming-specialized models, Flash-VStream, LiveStarPro, and TimeChat-Online. Maps are extracted at relative depths, resampled to 32×32, and min-max normalized. As Figure 4(a) shows, general-purpose models maintain sharp focus on micro-triggers across all layers, whereas streaming models progressively wash them into background noise, so the micro-trigger itself is lost.

This indicates that aggressive token merging in native streaming architectures systematically discards high-frequency, object-centric features. Future designs must preserve micro-spatial cues rather than prioritizing only global semantics under real-time compression.

5.3.2 Temporal Failure: Context Eviction and Goal Decay

Tracking query token attention weight and KV-Cache retention over a 60s stream in Figure 4(b), General-Purpose models lock query tokens via Prefix-Cache with only smooth attention dilution. In contrast, Streaming-Specialized models exhibit step-function cache collapses, each synchronized with a simultaneous plunge in query attention.



(a) (b)

Figure 4: **(a)** Averaged attention on Micro-triggers. Top: General-Purpose MLLMs. Bottom: Streaming-Specialized MLLMs. General-Purpose models keep sharp focus while streaming models wash triggers into the background. **(b)** KV-Cache retention and query attention over 60s. Top: General-Purpose MLLMs. Bottom: Streaming-Specialized MLLMs. General-Purpose models retain 100% with smooth decay, streaming models show step-function evictions causing synchronized attention collapse on the latent goal.

Thus, latent goal decay is driven by hard cache eviction rather than gradual attention dilution. This highlights the need for streaming memory mechanisms that explicitly preserve unfulfilled query tokens until the visual trigger appears.

5.3.3 Cascading Failure: Error Propagation in Multi-shot Scenarios

In Figure 5, second-trigger accuracy drops sharply following an initial error. For instance, Seed-1.6-vision's secondary failure rate jumps from 11.1% to 69.4%, and LiveStarPro hits 80.5%. To strictly isolate error propagation from inherent sample difficulty, we apply a context-correction intervention. Replacing the incorrect first answer with the ground truth improves second-trigger accuracy by 4.8 to 17.6 points on these failed samples (Appendix F.3).

These results effectively disentangle error propagation from sample difficulty. Since the intervention only corrects dialogue history while the video remains fixed, the recovered accuracy proves that early hallucinations actively contaminate ongoing context. Consequently, future streaming architectures urgently require mechanisms to correct early errors before they compound.

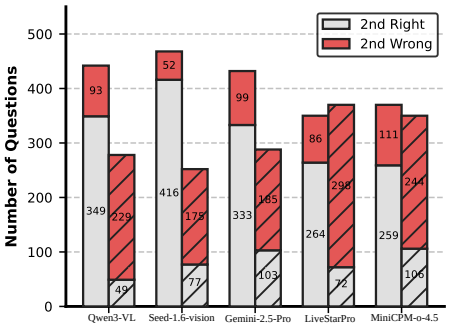


Figure 5: Error accumulation in Continuous Q&A across 1,440 samples. Left: Correct first answer. Right: Incorrect first answer.

6 Conclusion and Limitations

We introduce ASVBench, the first benchmark dedicated to evaluating asynchronous responses to fine-grained visual triggers in streaming video, comprising 1,452 videos and 4,350 high-quality Q&A pairs. Built upon a dual-axis framework spanning temporal and cognitive dimensions, ASVBench evaluates models across temporal complexities and trigger granularities. Extensive evaluations on 20 leading MLLMs reveal a substantial performance gap relative to human capabilities, with model accuracy degrading as asynchronous temporal tracking demands increase. Diagnostic analyses further expose four fundamental architectural bottlenecks: spatial over-smoothing of micro-triggers, temporal context eviction driving latent goal decay, impatience bias leading to premature hallucinations, and cascading error propagation in multi-shot interactions. These findings establish a diagnostic roadmap for reliable real-world streaming MLLMs.

Limitations: ASVBench currently utilizes frame-level visual triggers only. Region-level triggers and audio triggers remain important directions for future work.

REFERENCES

OpenAI:Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, FlorenciaLeoni Aleman, and Diogo Almeida. Gpt-4 technical report. Dec 2023.

Anthropic. Introducing claude 3.5 sonnet. https://www.anthropic.com/news/claude-3-5-sonnet, 2024. Accessed: 2026-07-28.

Anthropic. System card: Claude opus 4.6, 2026. URL https://www.anthropic.com/research/claude-opus-4-6. Accessed: 2026-07-27.

Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization, text reading, and beyond. *arXiv preprint arXiv:2308.12966*, 2023.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025.

ByteDance. Seed-1.6. https://seed.bytedance.com/en/seed1_6/, 2025.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. Temporalbench: Towards fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024. URL https://arxiv.org/abs/2410.10818.

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*, pp. 18407–18418, 2024.

Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. Livecc: Learning video llm with streaming speech transcription at scale. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2025.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, et al. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*, 2025.

Junbo Cui, Bokai Xu, Chongyi Wang, Tianyu Yu, Weiyue Sun, Yingjing Xu, Tianran Wang, Zhihui He, Wenshuo Ma, Tianchi Cai, et al. Minicpm-o 4.5: Towards real-time full-duplex omni-modal interaction, 2026. URL https://arxiv.org/abs/2604.27393.

Kaixuan Du, Xin Wan, YuKun Wang, Hang Zhang, Meng Cao, Dai Guan, Ming Chen, and Ni Li. Proactivebench: Can streaming video models really interact like humans?, 2026. URL https://arxiv.org/abs/2609.12658.

Lin Geng Foo, Jia Gong, Zhipeng Fan, and Jun Liu. System-status-aware adaptive network for online streaming video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 10514–10523, 2023.

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 *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 24108–24118, 2025.

GLM-5 Team. GLM-5: from vibe coding to agentic engineering, 2026. URL https://arxiv.org/abs/2602.15763.

Zhenpeng Huang, Xinhao Li, Jiaqi Li, Jing Wang, Xiangyu Zeng, Cheng Liang, Tao Wu, Xi Chen, Liang Li, and Limin Wang. Online video understanding: Ovbench and videochat-online. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 3328–3338, 2025.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024.

Kimi Team. Kimi K2.5: Visual agentic intelligence, 2026. URL https://arxiv.org/abs/2602.02276.

KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. May 2023.

Xinhao Li, Yuhan Zhu, Xiangyu Zeng, Yuhao Dong, Haoning Wu, Zhiqiu Zhang, Yuandong Yang, Changlian Ma, Qingyu Zhang, Yansong Shi, Xinyu Chen, Haoran Chen, Zizheng Huang, Jun Zhang, Kun Ouyang, Lin Sui, Ziang Yan, Yicheng Xu, Chenting Wang, Yinan He, Hongjie Zhang, Yi Wang, Yu Qiao, Yali Wang, Ziwei Liu, Kai Chen, and Limin Wang. VideoChat3: Fully open video MLLM for efficient and generalist video understanding, 2026. URL https://arxiv.org/abs/2607.14935.

Yifei Li, Junbo Niu, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, Pan Zhang, Yuhang Zang, Yuhang Cao, Conghui He, and Jiaqi Wang. OVO-Bench: How far is your video-LLMs from real-world online video understanding, 2025. URL https://arxiv.org/abs/2501.05510.

Junming Lin, Zheng Fang, Chi Chen, Haoxuan Cheng, Zihao Wan, Fuwen Luo, Ziyue Wang, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. In *ICASSP 2026-2026 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 12147–12151. IEEE, 2026.

OpenAI. GPT-5.5 system card, 2026. URL https://openai.com/index/gpt-5-5-system-card/. Accessed: 2026-07-27.

Jiawen Qian, Hang Du, Guoshun Nan, Shan Huang, Jiaqi Yu, Haoqing Wang, Jingfeng Chen, Mengxue Cai, Mengdi Yang, Jiarong Li, Zhilong Li, Hua wang, Jun Liu, Xudong Jiang, and Sicong Leng. Learning to respond: A large-scale benchmark and progressive learning framework for trigger-centric online video understanding, 2026. URL https://openreview.net/forum?id=gmpnSSiJt7.

Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360, 2024.

Yujiao Shen, Shulin Tian, Jingkang Yang, and Ziwei Liu. A simple baseline for streaming video understanding, 2026. URL https://arxiv.org/abs/2604.02317.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 18221–18232, 2024.

StepFun. Step-1o-turbo-vision. https://platform.stepfun.com/, 2025.

OpenBMB MiniCPM-o Team. Minicpm-o 2.6: A gpt-4o level mllm for vision, speech, and multimodal live streaming on your phone, 2025a.

Qwen Team. Qwen3-vl: Sharper vision, deeper thought, broader action. *Qwen Blog. Accessed*, pp. 10–04, 2025b.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Haomiao Xiong, Zongxin Yang, Jiazuo Yu, Yunzhi Zhuge, Lu Zhang, Jiawen Zhu, and Huchuan Lu. Streaming video understanding and multi-round interaction with memory-enhanced knowledge. *arXiv preprint arXiv:2501.13468*, 2025.

Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, and Jinzheng He. Qwen3-Omni technical report. *arXiv preprint arXiv:2509.17765*, 2025. URL https://arxiv.org/abs/2509.17765.

Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025a.

Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. SVBench: A benchmark with temporal multi-turn dialogues for streaming video understanding. *arXiv preprint arXiv:2502.10810*, 2025b. URL https://arxiv.org/abs/2502.10810. ICLR 2025.

Zhenyu Yang, Kairui Zhang, Bing Wang, Shengsheng Qian, and Changsheng Xu. LiveStarPro: Proactive streaming video understanding with hierarchical memory for long-horizon streams. *arXiv preprint arXiv:2606.17798*, 2026. URL https://arxiv.org/abs/2606.17798.

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.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, et al. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*, 2025.

Yuqian Yuan, Hang Zhang, Wentong Li, Zesen Cheng, Boqiang Zhang, Long Li, Xin Li, Deli Zhao, Wenqiao Zhang, Yueting Zhuang, et al. Videorefer suite: Advancing spatial-temporal object understanding with video llm. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pp. 18970–18980, 2025.

Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. Jun 2023.

Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, and Xiaojie Jin. Flash-vstream: Efficient real-time understanding for long video streams. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 21059–21069, 2025.

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*, 2026. URL https://openreview.net/forum?id=738HjJEbml.

Ruixiang Zhao, Jie Yang, Zijie Xin, Tianyi Wang, Fengyun Rao, Jing Lyu, and Xirong Li. Omnipro: A comprehensive benchmark for omni-proactive streaming video understanding, 2026. URL https://arxiv.org/abs/2605.18577.

Yucheng Zhao, Chong Luo, Chuanxin Tang, Dongdong Chen, Noel Codella, and Zheng-Jun Zha. Streaming video model. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 14602–14612, 2023.

A APPENDIX

B HUMAN EVALUATION ON ASVBENCH

This appendix provides the full verification protocols referenced in Step 7 of the main paper. It consists of two complementary stages: (i) an initial caption-level quality check on the data-engine outputs, and (ii) the final multi-stage Q&A verification with inter-annotator agreement statistics.

B.1 INITIAL CAPTION-LEVEL QUALITY CHECK

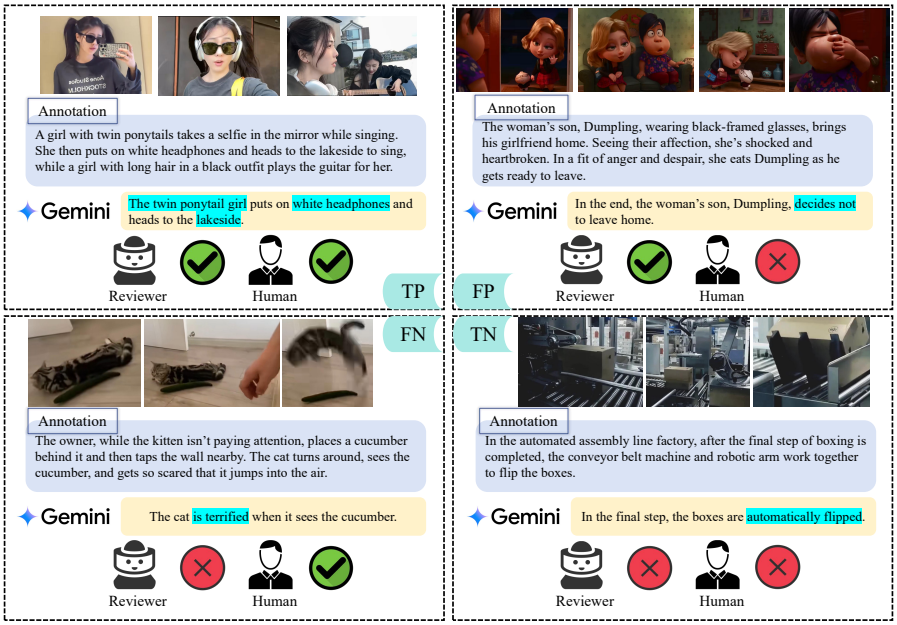


Figure 6: Initial human evaluation on the fine-grained information extraction of the data engine.

In the ASVBench framework, we first focus on assessing the accuracy and reliability of the fine-grained information extracted from video captions during the initial candidate generation phase. This evaluation process ensures that the raw pool of data is of high quality before the final filtering stages. To evaluate the correctness of the information, we conducted a manual review of the captions generated for a sample of video clips as shown in Figure 6.

The evaluation criteria for assessing the fine-grained information include several key metrics. True Positives indicate instances where the fine-grained details in the captions were correctly identified and confirmed as relevant after manual inspection. True Negatives represent instances where irrelevant or incorrect details were correctly excluded from the captions after review. False Positives denote instances where irrelevant or incorrect details were mistakenly included in the captions but were later found to be inaccurate. False Negatives signify instances where relevant details were omitted or overlooked in the captions but were identified as valid upon manual review.

To carry out this caption-level evaluation, we randomly selected 200 video segments from both the discarded and retained categories. The evaluation results are calculated as follows:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} = 0.81 \tag{6}$$

$$\text{Precision} = \frac{TP}{TP + FP} = 0.89 \tag{7}$$

$$\text{Recall} = \frac{TP}{TP + FN} = 0.76 \tag{8}$$

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} = 0.82 \tag{9}$$

The precision score of 0.89 indicates that the majority of fine-grained details identified in the initial captions are indeed accurate. This demonstrates the effectiveness of our dataset creation process in providing reliable candidates for the subsequent Q&A verification stage.

B.2 Final Q&A Verification Protocol

As described in Step 7 of the main text, professional annotators and Qwen3-Omni perform multi-stage verification, filtering 5,600 initial Q&A pairs down to 4,350 (22.32% rejection rate). To quantify annotation reliability, three professional annotators independently re-label a random sample of 700 Q&A pairs, yielding a Fleiss' Kappa of 0.91, which indicates near-perfect agreement. Separately, the Human Baseline on ASVBench reaches an overall accuracy of 95.47% (Table 2), confirming that the retained samples are objective and grounded in human perception. Among the 1,250 rejected pairs, Visual Ambiguity accounts for 57% (triggers too small or motion-blurred), Subjective Interpretation for 21% (no absolute ground truth), and LLM Hallucination for 22% (fabricated or mis-causal content). Together, the initial caption-level check and the final Q&A verification protocol significantly contribute to the overall reliability and accuracy of the ASVBench dataset.

C More Details of Dataset

C.1 Question Format and Chance Level

Of the 4,350 questions, 3,140 are four-option single-choice items and 1,210 are true/false items. A uniform guess is correct with probability 25% on a single-choice item and 50% on a true/false item, so the expected number of correct answers is

$$3{,}140 \times 0.25 + 1{,}210 \times 0.50 = 785 + 605 = 1{,}390. \tag{10}$$

The expected accuracy is $1{,}390/4{,}350 \approx 31.95\%$. The text-only GPT-4o result in Table 2 is 30.64% overall, closely approximating this chance level.

C.2 Sample Distribution of Dataset

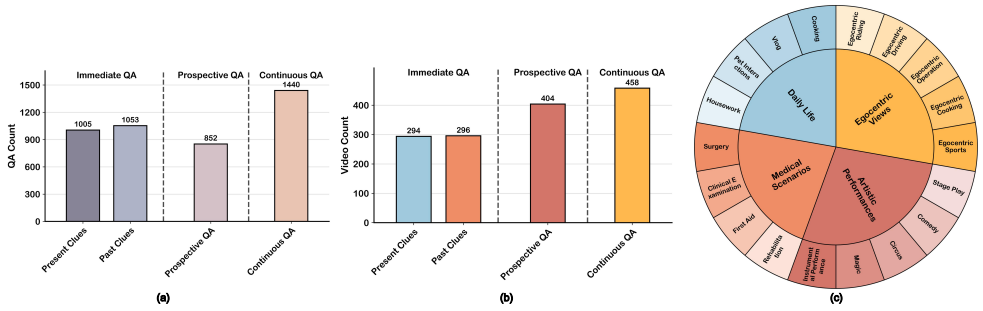


Figure 7: Sample distribution of the dataset showing the query count by clue type, video sample distribution by query type, and the distribution across various video categories.

Figure 7 presents a detailed distribution of the samples in the dataset which is crucial for understanding the composition and structure of the data used in our experiments. The dataset is organized into three primary types of tasks with each presenting distinct characteristics and challenges.

C.2.1 Distribution of Q&A Types

The first category is Immediate Q&A requiring the model to provide instant responses based on available clues. This category is further divided into two types namely Present Clues and Past Clues. The Present Clues consist of 1005 entries directly available in the current context. The Past Clues contain 1053 entries referring to clues revealed previously. In total, Immediate Q&A consists of 2058 entries across the two clue types.

The second category is Prospective Q&A involving questions requiring the model to answer based on fine grained future events. There are 852 entries in this category.

Table 3: Query examples for different tasks in ASVBench.

Task	Subtasks	Query Examples
Immediate Q&A Clues in the present	Recognition	What is the jersey number of the baseball player from Mexico signing autographs at 00:04:09
	Comprehension	Based on the actions of the people when the black dog falls into the pool, what is the most appropriate description of the situation
	Reasoning	The group had initially planned for an outdoor hotpot by Sayram Lake. Based on the video context at 14:34, what was the most likely reason for them to move their hotpot meal indoors
Immediate Q&A Clues in the past	Recognition	Before the magician successfully selected an audience member for the card trick, how many times was the orange ball caught by someone who was then rejected by the audience
	Comprehension	Did Raju get into the white Toyota Innova before Farhan did
	Reasoning	Based on the continuous activities with the laptop, keyboard, and notebooks between 00:04:20 and 00:05:30, what was the primary focus of the person
Prospective Q&A	Recognition	When the video shows female players for comparison, what is the jersey color of Sun Yingsha
	Comprehension	Monitor the grey cat wearing the door stopper collar. Does it successfully remove the collar by itself
	Reasoning	Based on the statement of the doctor, infer why the girl could only see after the specific eye drops were administered
Continuous Q&A	Recognition	What are the measured weights of the 3 star straight and shakehand paddles
	Comprehension	According to the text overlays, what are the initial strategies mentioned by Simon Gauzy, and how does he adapt his strategy later in the match
	Reasoning	Why does the magician ask the female assistant to sit on the opposite side of the table before performing the Oreo cookie restoration trick

Finally, Continuous Q&A requires the model to provide ongoing responses related to future events continuing over time rather than being confined to a single instance. This task involves 1440 entries testing the capability of the model to track and respond to dynamic and evolving scenarios.

C.2.2 Video Segment Distribution

Figure 7 provides an overview of the number of video segments corresponding to each task type. In terms of video segments, Immediate Q&A includes 294 segments with Present Clues and 296 segments with Past Clues. Prospective Q&A is associated with 404 video segments. Continuous Q&A involves 458 segments in the collection.

C.2.3 Dataset Categories

Figure 7 illustrates the broad categorization of the dataset covering a range of domains commonly encountered in daily life. The major categories include Daily Life, Medical Scenarios, Artistic Performances, and Egocentric Views. These categories capture a wide array of real world scenarios, making the dataset highly relevant for various practical applications.

Each of these broad categories is further subdivided into specific subcategories to provide a more granular structure. Specifically, Daily Life includes Cooking, Vlog, Pet Interactions, and House-

work; Medical Scenarios covers Surgery, Clinical Examination, First Aid, and Rehabilitation; Artistic Performances comprises Instrumental Performance, Magic, Circus, Comedy, and Stage Play; while Egocentric Views features Egocentric Sports, Egocentric Cooking, Egocentric Operation, Egocentric Driving, and Egocentric Riding. This hierarchical structure allows for a more detailed analysis of model performance across different domains and provides valuable insights into how well a model generalizes across diverse contexts.

C.3 Task Example and Data Format

In this section, we provide a breakdown of various task examples from the dataset focusing on different formats and their associated tasks. The dataset includes multiple types such as Immediate Q&A, Prospective Q&A, and Continuous Q&A. Each involves distinct subtasks like recognition, comprehension, and reasoning. Below we explore the task examples corresponding to these formats as shown in Table 3 of this appendix section.

Immediate Q&A (Clues in the present)

"Question Type": "Single Choice",
"Question Task": "Fine-grained Reasoning",
"Detail Task": "Physical State Reasoning",
"Question": "Based on the visible steam at night, what is the temperature of the outdoor onsen water?",
"Timestamp(begin question)": "00:07:29",
"Timestamp(answer question)": "00:07:29",
"Options":
"A": "Hot",
"B": "Cold",
"C": "Lukewarm",
"D": "Undetermined"

"Correct Answer": "A"

Immediate Q&A (Clues in the past)

"Question Type": "Single Choice",
"Question Task": "Fine-grained Reasoning",
"Detail Task": "Causal Chain Retrospection",
"Question": "According to the aerodynamic explanation for the hollow frisbee, what causes the increased impact lift?",
"Timestamp(event_start)": "00:02:14.00",
"Timestamp(event_end)": "00:02:22.00",
"Timestamp(begin question)": "00:02:25",
"Timestamp(answer question)": "00:02:27",
"Options":
"A": "Rapid self-rotation providing gyroscopic stability.",
"B": "Airflow clinging to the curved surface causing pressure differential.",
"C": "The formation of three separation bubbles forcing airflow downwards.",
"D": "The frisbee being solid and not hollow."
"Correct Answer": "C"

Prospective Q&A

"Question Type": "True/False",
"Question Task": "Fine-grained Comprehension",
"Detail Task": "Interaction Watch",
"Question": "Will any audience members be invited onto the stage to physically interact with the magic trick's props?",
"Timestamp(begin question)": "00:00:50",
"Timestamp(event_start)": "00:01:26",
"Timestamp(answer question)": "00:01:34",
"Timestamp(event_end)": "00:01:46",
"Options":
"A": "YES",
"B": "NO"

"Correct Answer": "A"

Continuous Q&A

"Question Type": "Single Choice",
"Question Task": "Fine-grained Recognition",
"Detail Task": "Specific Attribute Trigger",
"Question": "Observe Zhang Xiaofan/Gui Li's clothing. What is the main color of his attire in two different appearances?",
"Timestamp(event_start)": "00:04:03",
"Timestamp(event_end)": "00:05:12",
"Timestamp(begin question)": "00:04:00",
"Timestamp(answer question time1)": "00:04:05",
"Timestamp(answer question time2)": "00:05:08",
"Options":
"A": "White and blue",
"B": "Green",
"C": "Dark (black/red)",
"D": "Golden"
"time1_Correct Answer": "A",
"time2_Correct Answer": "C"

Figure 8: Data format examples for different tasks

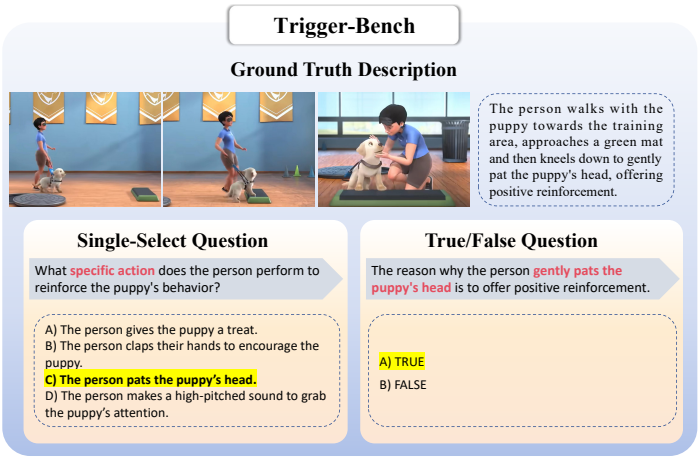


Figure 9: Visualization of ASVBench Q&A example. For the convenience of testing, our questions are set as Single Select Questions and True False Questions.

For Immediate Q&A, two types of clues are used including those derived from the present context and those rooted in past clues from earlier frames.

Immediate Q&A with Clues in the Present typically involves questions related to real time observation and comprehension. For example, recognition tasks may ask for a simple identification like the jersey number of a baseball player signing autographs at a specific timestamp. Comprehension tasks dive deeper into the video context such as describing the situation when a dog falls into a pool. Reasoning tasks require a higher level of inference like determining the most probable reason a group moved their outdoor meal indoors based on contextual clues.

Immediate Q&A with Clues in the Past involves questions requiring a retrospective look at past events in the video. Recognition tasks may ask about past actions such as how many times a ball was caught by a person before being rejected. Comprehension tasks might ask if a character enters a car before another person. Reasoning tasks involve more complex inferences such as determining the primary focus of a person based on their continuous interaction with objects during a specific time period that the question refers to.

For Prospective Q&A, the questions are generally forward looking and rely on understanding based on earlier video events. Recognition tasks might focus on identifying the jersey color of a player when female players are shown in comparison. Comprehension tasks might involve monitoring a cat wearing a collar to see if it successfully removes the collar. Reasoning tasks may require inferring the cause of a specific action based on a previous explanation.

Lastly, Continuous Q&A focuses on questions requiring ongoing observation over a period of time with a focus on changes or progressions. Recognition tasks might ask for the measured weights of specific items. Comprehension tasks could ask for details about how the strategies of a player evolve during a match based on text overlays shown in the video. Reasoning tasks require deeper inference such as explaining why a magician asks an assistant to sit on a specific side of the table during a trick performed in view of the camera.

These examples illustrate the diversity of task types in the dataset emphasizing the range of cognitive and observational skills required for effective streaming video understanding. This diversity ensures that the dataset provides a comprehensive evaluation of multimodal systems testing their performance in recognizing, comprehending, and reasoning about various aspects of the visual information present in the video stream.

D Ablation Study on Causal Sliding Window Size

In the main text, we adopt a causal sliding window strategy to evaluate offline general-purpose models under strict streaming constraints. To balance visual perception fidelity and real-time computational overhead, we empirically set the window size Δw to 32 frames. Here, we provide the detailed ablation study justifying this architectural choice.

To ensure evaluation efficiency, we conduct this ablation study on a representative subset of 1,540 randomly sampled Q&A pairs from the ASVBench dataset, so the accuracies below are not the full-set scores in Table 2. We compare $\Delta w = 32$ against $\Delta w = 64$ on three open-source models whose KV-cache and step latency can be measured directly. As shown in Table 4, expanding the window to 64 frames yields only a marginal gain: $+1.23$ for Qwen3-VL (8B), $+1.35$ for MiniCPM-V-4.5 (9B), and $+1.47$ for Qwen3-VL (30B-A3B).

However, this minor gain comes at a prohibitive cost. Because self-attention scales quadratically with sequence length, doubling the frame count inflates the KV-cache. Qwen3-VL (30B-A3B) rises from 32.4 GB to 85.6 GB ($+164\%$), which exceeds a standard consumer GPU, and its step latency rises from 0.95s to 2.68s ($+182\%$). The two smaller models show the same pattern, with latency increases of $+160\%$ to $+173\%$. These overheads violate the real-time constraint of streaming interaction with a user waiting for the trigger.

Therefore, setting $\Delta w = 32$ serves as the optimal trade-off, successfully decoupling visual perception capabilities from architectural flaws while strictly adhering to real-world latency boundaries.

E More Benchmark Visualization

In this section, we provide a visualization of the question answering process within the ASVBench framework as depicted in Figure 9. For each processed video segment, we generate captions and

Table 4: Ablation study on the causal sliding window size (Δw).

Model	Δw (Frames)	Accuracy (%)	KV-Cache (GB)	Latency (s/step)
Qwen3-VL (8B)	32	64.92	14.2	0.42
	64	66.15 (+1.23)	35.8 (+152%)	1.15 (+173%)
MiniCPM-V-4.5 (9B)	32	68.10	15.6	0.48
	64	69.45 (+1.35)	38.5 (+146%)	1.25 (+160%)
Qwen3-VL (30B-A3B)	32	66.85	32.4	0.95
	64	68.32 (+1.47)	85.6 (+164%)	2.68 (+182%)



Figure 10: Visualization of ASVBench data across different question answering timelines.

design questions that focus on fine grained distinctions between objects and events with answers being objective and clear. To ensure a more comprehensive benchmark, we also incorporate the concept of asynchronous responses shown in Figure 10. This is represented by questions that not only receive immediate answers but also account for future and continuous responses. The question answering process is thus organized into three distinct parts including Current Question and Current Answer, Current Question and Future Answer, and Current Question and Continuous Answer. This structure helps evaluate the ability of the system to handle varying response dynamics and improves the overall robustness of the benchmark.

F More Experimental Results

F.1 Impact of Temporal Spacing

To further investigate the influence of temporal factors on model performance, we analyze the effect of temporal spacing. This spacing represents the time interval between the moment a question is asked and the moment the answer occurs in the video. As shown in Figure 11, the results of five models reveal a clear trend where a longer temporal spacing leads to lower accuracy.

This phenomenon can be attributed to the increased difficulty in maintaining the latent goal of the query over extended periods. This indicates that models should enhance their ability to memorize and understand context. Such enhancement enables them to adapt to the preservation of latent intent over long time horizons in asynchronous streaming video understanding.

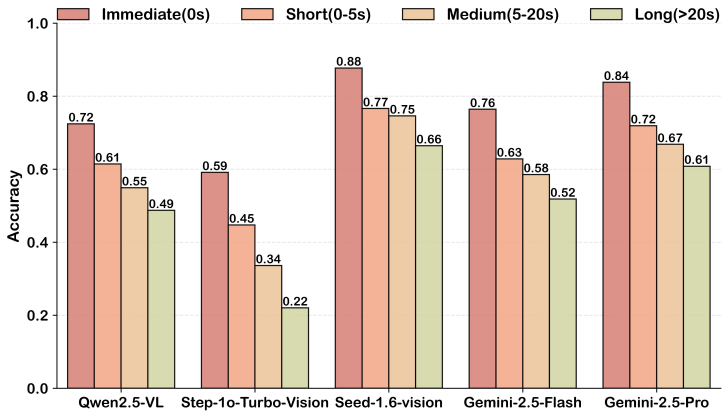


Figure 11: **Performance of temporal spacing.** We select five prominent models to demonstrate the impact of temporal spacing on model accuracy.

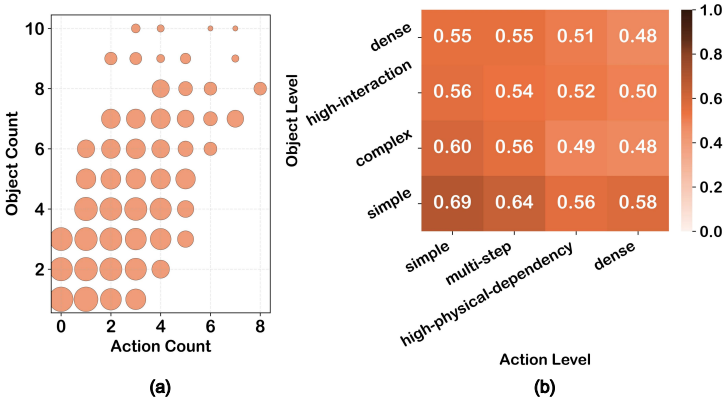


Figure 12: **The impact of video objects and actions.** (a) Average model accuracy by object/action quantities. Circle radius denotes accuracy, where larger means higher; (b) Average model accuracy by object/action types. Color shade denotes accuracy, where darker means higher.

F.2 Impact of Object and Action Complexity

To further explore the impact of object and action complexity on model performance, we conduct a fine-grained analysis by annotating the number of objects and actions involved in each question. Additionally, we categorize the types of objects and actions based on the spatiotemporal characteristics of each annotated video event.

As shown in Figure 12, we observe a general trend: models experience performance degradation as the number and complexity of objects and actions increase. This trend can be attributed to the increased difficulty in parsing and reasoning about multiple entities and their interactions, which highlights the pressing need for models to enhance their capabilities in handling complex scenes for improved fine-grained understanding. Detailed analyses regarding specific object/action counts and types are presented in the following subsections.

F.2.1 Accuracy Across Varying Object and Action Counts

In this section, we evaluate the performance of various models in understanding streaming videos, specifically focusing on their capacity to process multiple objects and actions simultaneously. Overall, current models exhibit limited proficiency in this aspect. On average, models can reliably focus on only three to four fine-grained objects and interacting actions at a time.

Gemini-2.5-Pro achieves the best performance among all evaluated models, demonstrating the capability to simultaneously handle six to seven different objects and actions. Notably, it exhibits a stronger aptitude for perceiving actions compared to objects, with a particular emphasis on interactions. Remarkably, it can manage up to eight types of interactions with impressive accuracy.

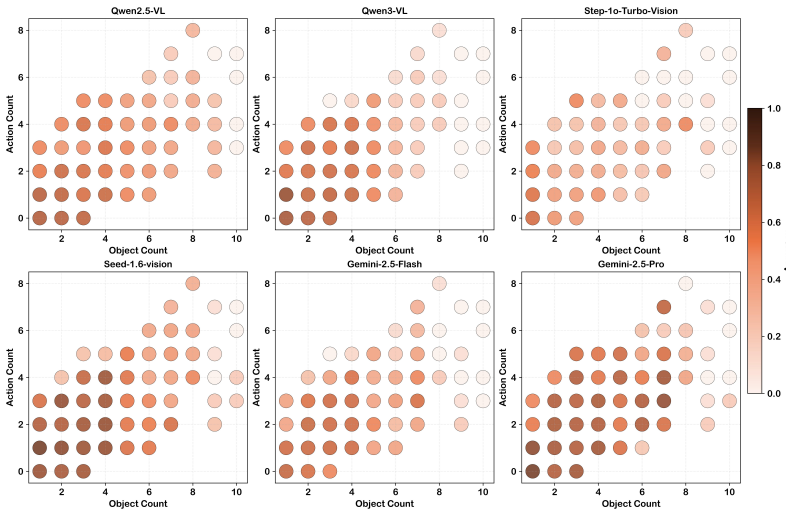


Figure 13: **The accuracy of representative models across different object and action quantities.** The shade of color indicates the average accuracy, with darker colors representing higher accuracy.

Following closely is Seed-1.6-vision, which can effectively process four to five distinct objects and actions concurrently. Similar to Gemini-2.5-Pro, Seed-1.6-vision shows a greater proficiency in understanding actions rather than isolated objects.

Qwen3-VL and Gemini-2.5-Flash demonstrate moderate capabilities, processing three to four objects and actions simultaneously. While their performance is constrained compared to the top-tier models, they still exhibit reasonable competency in handling fine-grained interactions. Conversely, Qwen2.5-VL and Step-1o-Turbo-Vision exhibit more limited capacities. Qwen2.5-VL handles only two to three objects and actions, while Step-1o-Turbo-Vision struggles to process more than two. Interestingly, Step-1o-Turbo-Vision excels in perceiving objects over actions, showing a stronger bias toward object detection rather than interaction modeling.

In summary, while model capabilities vary significantly, a prevailing trend indicates that current models tend to prioritize objects over actions, as physical entities offer higher visual stability for perception. Although interaction modeling remains a distinct strength for top-performing models, the overall capacity to capture complex actions is still secondary to object recognition. Nevertheless, simultaneously handling multiple objects and actions in streaming video scenarios remains a universal challenge for current streaming models.

F.2.2 Accuracy Across Different Object and Action Types

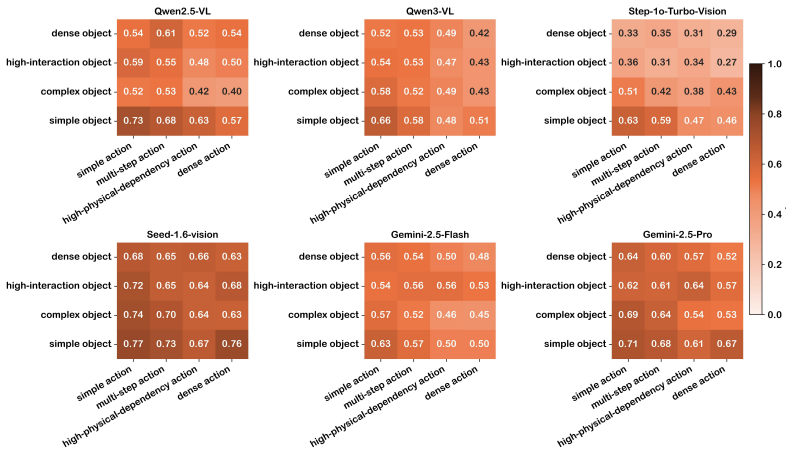


Figure 14: **The accuracy of representative models over object and action types.** The shade of color indicates the average accuracy, with darker colors representing higher performance.

To comprehensively assess model performance on diverse fine-grained objects and actions, we categorize them into four distinct groups based on their spatiotemporal relationships. Objects are classified as *simple*, *complex*, *highly interactive*, and *dense*. Similarly, actions are categorized as *simple*, *multi-step*, *highly physically dependent*, and *dense*.

Overall, current models yield relatively low accuracy across these fine-grained categories, indicating that their capacity for streaming video understanding remains insufficient, particularly for dense and highly interactive tasks. Interestingly, we observe that when the action is fixed as a "simple action", accuracy decreases sharply as object complexity increases. In contrast, when the object is fixed as a "simple object", accuracy experiences a milder decline as action complexity increases. This suggests that models struggle more with complex objects than with complex actions, indicating that capturing complex objects is a more formidable challenge for current architectures.

Seed-1.6-vision achieves the highest overall accuracy, with all scores exceeding 60%, though it shows slight vulnerability to dense objects and complex actions. Gemini-2.5-Pro ranks second, maintaining scores above 50% across all categories. Gemini-2.5-Flash follows closely, sustaining scores above 45% and exhibiting performance degradation trends similar to the aforementioned models as task complexity rises.

Qwen2.5-VL maintains scores above 40%, peaking at 73% for the combination of simple objects and simple actions. While it demonstrates a robust understanding of simple entities and movements, its performance significantly drops on complex objects, revealing a noticeable gap between simple and complex object comprehension. Qwen3-VL similarly performs well on simple object-action combinations but struggles considerably with dense actions and videos featuring rich motion dynamics. Finally, Step-1o-Turbo-Vision records the lowest performance across all object categories, indicating a limited ability to comprehend objects in diverse scenarios.

In conclusion, while models like Seed-1.6-vision and Gemini-2.5-Pro demonstrate robust baseline performances, significant challenges persist across varying object and action types. This underscores the necessity for future model improvements to effectively handle highly complex scenarios in streaming video understanding.

Table 5: Impact of context correction on second-trigger accuracy for initial failure samples.

Model	**Contaminated (%)**	**Corrected (%)**	**Gain**
Qwen3-VL (8B)	17.6	35.2	+17.6
Seed-1.6-vision	30.6	46.8	+16.2
Gemini-2.5-Pro	35.8	43.5	+7.7
LiveStarPro (8B)	19.5	31.4	+11.9
MiniCPM-o-4.5 (9B)	30.3	35.1	+4.8

F.3 Intervention Analysis: Error Propagation versus Sample Difficulty

As discussed in Section 5.3.3, first-trigger and second-trigger failures in Continuous Q&A are positively associated. That association confounds error propagation with intrinsic sample difficulty: items that are missed once may simply be harder. To separate the two, we run a context-correction intervention, equivalent to injecting the ground-truth first answer.

We use the five models in Figure 5: Qwen3-VL (8B), Seed-1.6-vision, Gemini-2.5-Pro, LiveStarPro (8B), and MiniCPM-o-4.5 (9B). We keep only Continuous Q&A samples on which the model's first answer is wrong. In the standard setting, the dialogue history still contains that hallucinated answer when the model answers the second trigger. In the corrected setting, we replace the wrong first answer with the ground truth before the second response, so the model reasons from an uncontaminated history of that sample's first answer.

Table 5 reports the comparison. After the history is corrected, second-trigger accuracy rises for every model, with absolute gains from $+4.8$ to $+17.6$ points. Because the video is unchanged and only the previous answer is repaired, the gain is evidence that the hallucinated first state actively contaminates later reasoning, rather than merely marking a harder item. Error propagation is therefore a catalyst of the cascade, and continuous streaming architectures need a way to correct an early mistake.

G AI USE STATEMENT

In this work, we used generative AI tools for dataset construction and quality control. Specifically, GPT-4o was used to extract object- and time-attention annotations from source videos; Claude-3.5-Sonnet was used to assign task categories; Gemini-2.5-Pro was used to generate timestamped Q&A pairs; and Qwen-VL-Max and Qwen3-Omni were used to filter and verify annotations. Professional human annotators independently verified a subset of the data, and all AI-generated annotations were subject to multi-stage human review. We have reviewed all AI-assisted work, including generated questions, answers, timestamps, and filtering decisions. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

H REPRODUCIBILITY STATEMENT

We provide dataset statistics, the dual-axis task formulation, and both streaming and causal sliding-window inference protocols in the main paper. Annotation pipeline details are described in Section 4.1 and Section 4.2, with additional verification protocols in Appendix B, domain statistics in Appendix C, and window-size ablations in Appendix D. Upon acceptance, we will release ASVBench, including videos, timestamped Q&A pairs, evaluation scripts, and the sliding-window protocol, to facilitate reproduction of the reported results.