FASTBENCH: CAN STREAMING VLMS PERCEIVE HIGH-DYNAMIC REAL-WORLD STREAMS?

Anonymous authors
Paper under double-blind review

ABSTRACT

Streaming Video Large Language Models (VLMs) have emerged as a promising paradigm for real-time, continuous video understanding. However, existing streaming benchmarks predominantly target low-dynamic scenarios, masking a critical challenge: under a bounded context budget, a model must trade off long temporal history, spatial resolution, and temporal granularity, and the sparse sampling rates (1–2 FPS) adopted by current systems inevitably miss fast-paced events. To bridge this gap, we introduce FastBench, a benchmark designed to evaluate streaming VLMs on high-dynamic real-world video streams. FastBench is built with a trajectory-grounded pipeline: a VLM proposes candidate QA pairs from native high-frame-rate clips, a low-FPS filter discards pseudo-dynamic questions that remain answerable at 2 FPS, and a Vision Expert Verification stage grounds answers on object trajectories extracted by SAM3 and CoTracker3 to suppress hallucinations, followed by three rounds of human inspection. The resulting 306 QA pairs span eight domains (e.g., sports, gaming, wildlife, and transportation), six capability dimensions, and forward, instant, and backward temporal scopes, each with human-annotated evidence intervals. Alongside the benchmark, we present ProactiveFrame, a training-free baseline that lets a model raise or restore the frame rate of incoming chunks through text tokens, supported by a dual-tier sliding window that buffers recent high-FPS chunks and gracefully degrades them into sparse history. Experiments show that high-dynamic perception remains far from solved: the strongest model, Gemini-3.5-Flash, reaches only 50.7%. Denser sampling helps substantially (Qwen3-VL-8B improves from 32.9% at 2 FPS to 44.6% at 24 FPS) but saturates as history is compressed. ProactiveFrame improves over sparse uniform sampling (+5.4% and +1.5%), yet stays well below oracle-guided focusing, revealing that current VLMs struggle to decide, from the visual stream alone, when finer temporal perception is needed. FastBench establishes a rigorous testbed and offers actionable insights for high-dynamic streaming video understanding.

1 INTRODUCTION

The remarkable success of Large Language Models (LLMs) and Vision-Language Models (VLMs) has catalyzed significant progress in the domain of video understanding. Recent advancements in Video VLMs, such as VideoChat (Li et al., 2025a) and LLaVA-NeXT-Video (Zhang et al., 2024), have demonstrated impressive capabilities in comprehending temporally bounded, offline video clips. However, real-world applications often require interacting with continuous, infinite visual inputs on the fly. This necessity has driven a recent paradigm shift towards Streaming Video Large Language Models (Streaming VLMs), which are designed to process newly arriving visual streams in real time and maintain long-term context, as exemplified by StreamingVLM (Xu et al., 2026).

Despite this rapid progress, existing streaming evaluation benchmarks, such as OVO-Bench (Niu et al., 2025) and PhoStream (Lu et al., 2026b), emphasize temporal reasoning and response timing without explicitly targeting high-dynamic perception. This leaves a critical challenge encountered in real-world streaming applications: accurately perceiving and comprehending highly dynamic, fast-paced events under strict context window constraints. In practice, video streams exhibit a heterogeneous mixture of low- and high-dynamic scenes. Operating under limited token budgets, current models must balance the need for maintaining a long temporal history, preserving high spatial resolution, and capturing fine-grained temporal information. Such constraints motivate sparse frame

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.

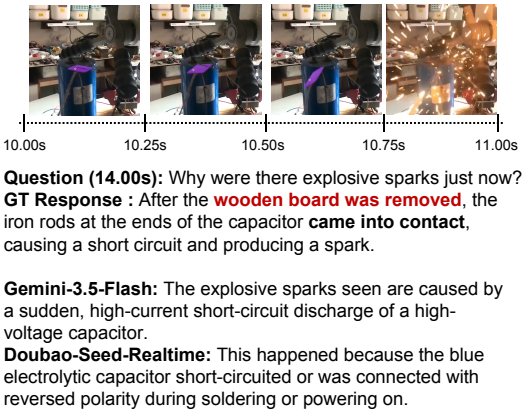


Figure 1: **A perception failure in a high-dynamic video event.** A wooden separator (highlighted in purple using SAM 3) is removed, allowing the metal rods to touch and produce sparks. Both models provide generic electrical explanations that could be inferred from the sparks alone, but fail to identify the brief sequence of separator removal and subsequent contact that triggers the discharge. This illustrates the challenge of capturing critical visual details in fast-paced events.

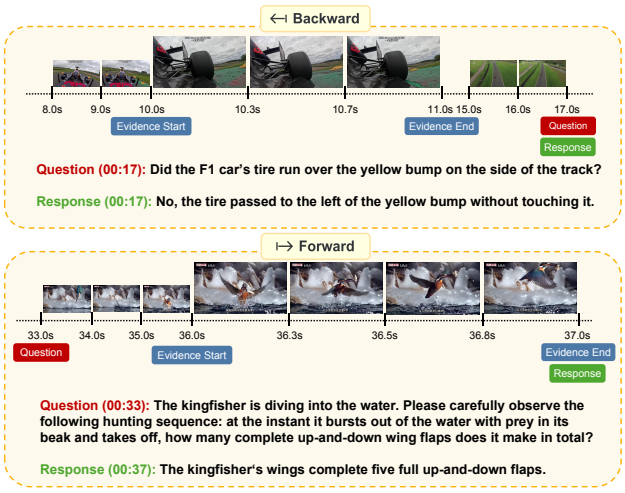
sampling; for example, the online baselines evaluated in OVO-Bench operate at 1–2 frames per second (Niu et al., 2025). Consequently, when faced with rapidly changing high-frequency actions—such as sports, extreme gaming, or wildlife hunting—these models often miss crucial visual details, leading to severe perception failures (as illustrated in Fig. 1). While recent studies have introduced offline benchmarks focusing on motion comprehension and high frame-per-second (FPS) perception (e.g., MotionBench, Hong et al., 2025; FPS-Bench, Choudhury et al., 2026), evaluating the capability to perceive high-dynamic events within a continuous, untrimmed streaming context remains an open problem.

To bridge this critical gap, we introduce FastBench, a novel benchmark specifically designed to evaluate the performance of streaming VLMs on high-dynamic real-world video streams. FastBench is constructed via a meticulously designed pipeline that synergizes VLM-driven generation, agentic expert verification, and rigorous human inspection. To ensure true temporal dependency, we introduce a low-FPS verification step that proactively filters out pseudo-dynamic questions. Furthermore, to overcome the inherent hallucination risks of VLMs in fast-paced scenarios, we propose a novel Vision Expert Verification mechanism. By autonomously invoking vision foundation models, the VLM verifies and refines its QA pairs against objective, text-based physical trajectories rather than error-prone visual features. Spanning diverse domains including sports, gaming, nature, and traffic, FastBench establishes a stringent evaluation standard for high-dynamic streaming perception.

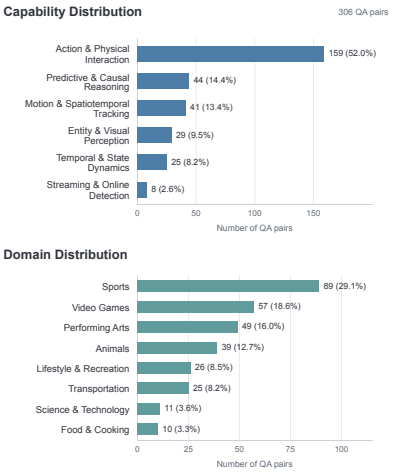
Along with the benchmark, we present ProactiveFrame, a simple, training-free baseline framework. Diverging from passive uniform sampling, ProactiveFrame empowers the agent to dynamically adjust its temporal perception granularity via text prompts based on the ongoing streaming context. To balance the need for fine-grained perception and long-term memory under strict token budgets, we introduce a dual-tier context sliding window mechanism. Operating on a single, unified chronological stream, this design proactively buffers high-FPS chunks for fleeting events, which subsequently undergo a graceful temporal degradation into a sparse historical context. This hierarchical management elegantly resolves the trade-off, enabling continuous high-dynamic perception without fragmenting the temporal horizon. Evaluations on FastBench show that high-dynamic perception benefits from denser temporal sampling, but remains constrained by the trade-off between temporal granularity and historical context. ProactiveFrame improves over sparse uniform sampling, yet its gap to oracle-guided sampling highlights the challenge of autonomously identifying when finer temporal perception is needed.

In summary, our main contributions are three-fold:

- **High-Dynamic Streaming Benchmark:** We introduce **FastBench**, a benchmark specifically designed to evaluate high-frame-rate-dependent perception in continuous video streams under bounded-context and response-timing constraints



(a) **Temporal QA examples.** The backward query at 17 s refers to evidence at 10–11 s; the forward query at 33 s is answered after observing evidence at 36–37 s.



(b) **Dataset distributions.** Counts and percentages of 306 QA pairs across six capability categories (top) and eight domains (bottom).

Figure 2: **Overview of FastBench.** Evidence Start and Evidence End in (a) mark the human-annotated interval containing the visual evidence required to answer each question.

- **Trajectory-Grounded Data Pipeline:** We propose an innovative data construction pipeline incorporating a low-FPS filtering mechanism to remove pseudo-dynamic questions, paired with a *Vision Expert Verification* module that grounds VLM annotations on objective physical trajectories to mitigate visual hallucinations.
- **Training-Free Baseline Framework:** We present **ProactiveFrame**, a simple yet effective training-free framework that features a dual-tier context sliding window mechanism. It dynamically buffers high-FPS chunks for fleeting events and allows them to gracefully degrade into sparse historical context, balancing fine-grained perception and long-term memory under tight token budgets.

2 Related Work

2.1 Streaming Video Understanding Benchmarks

Streaming video benchmarks evaluate understanding from incrementally available evidence. StreamingBench (Lin et al., 2026) assesses real-time visual, omni-source, and contextual understanding, while OVO-Bench (Niu et al., 2025) covers backward tracing, real-time understanding, and forward active responding. OVBench (Huang et al., 2025) evaluates online perception, memory, and reasoning across past, present, and future contexts. PhoStream (Lu et al., 2026b) extends evaluation to open-ended question answering in mobile scenarios, emphasizing audio-visual evidence and response timing. Building on these temporal task structures, FastBench focuses on high-frame-rate-dependent evidence in continuous streams, jointly evaluating fine-grained perception, historical retention, and response timing under a bounded visual context.

2.2 Fine-Grained Motion and High-Frame-Rate Understanding

Fine-grained temporal benchmarks seek to distinguish motion understanding from reliance on static cues. TempCompass (Liu et al., 2024) uses videos with shared static content but contrasting temporal properties. MotionBench (Hong et al., 2025) evaluates fine-grained motion comprehension and investigates higher-frame-rate inputs and motion-preserving feature fusion. FPS-Bench (Choudhury et al., 2026) introduces minimum frames-per-second (minFPS), using human annotations to identify the lowest sampling rate needed to verify an answer. Complementing these efforts, FastBench uses

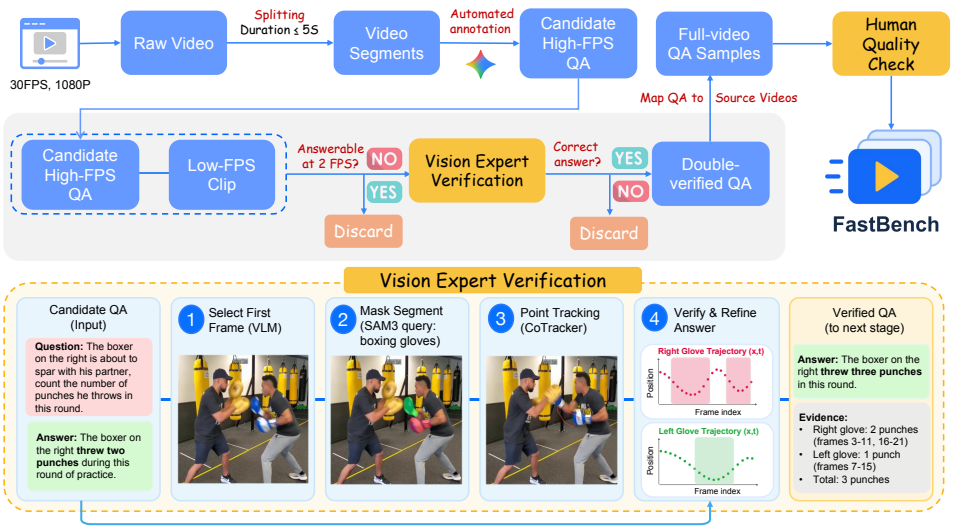


Figure 3: **FastBench data construction pipeline.** A VLM generates candidate QA pairs from short, high-frame-rate video clips. Questions answerable at 2 FPS are discarded. The remaining pairs undergo automated vision-expert verification, where the VLM uses SAM3 segmentation and CoTracker3 trajectories to verify or refine answers; pairs that cannot be verified or corrected are discarded. Retained QA pairs are mapped back to their source videos and undergo human quality checks to form FastBench. The lower panel illustrates how trajectory evidence supports correcting a punch count from two to three.

low-FPS answerability filtering rather than per-question minFPS estimation to evaluate high-frame-rate-dependent evidence within a causal streaming protocol.

2.3 Proactive Perception and Streaming Context Management

Streaming systems must coordinate response timing, visual computation, and context retention. VideoLLM-online (Chen et al., 2024) learns from video streams to support temporally aligned dialogue. LION-FS (Li et al., 2025b) separates fast response determination from detail-enhanced generation, while StreamMind (Ding et al., 2025) uses event-gated LLM invocation to decouple perception from language reasoning. VideoChat3 (Li et al., 2026) adapts the spatial resolution of subsequent windows using predicted interaction states. For context management, StreamingVLM (Xu et al., 2026) maintains different KV-cache retention windows for visual and textual tokens, while simpleStream (Shen et al., 2026) demonstrates a strong recent-frame baseline and a perception–memory trade-off. In contrast, ProactiveFrame provides a training-free, prompt-driven baseline that adjusts future temporal sampling rates and uses a dual-tier window to retain recent high-FPS observations alongside sparse historical context.

3 FastBench

In this section, we present the design of FastBench and our baseline framework, ProactiveFrame. We first detail the data curation pipeline in Section 3.1 and present the dataset statistics in Section 3.2. Finally, Section 3.3 describes the framework and core dynamic perception mechanisms of ProactiveFrame.

3.1 Data Curation

Raw Video Collection. To construct a diverse pool of candidate videos rich in high-dynamic, fast-paced events across domains such as sports, gaming, nature, and traffic, we introduce an automated, metadata-driven collection pipeline. We first employ a VLM, Gemini-3.1-Pro (Google DeepMind, 2026a), to brainstorm and generate a comprehensive set of search keywords corresponding to our target high-dynamic scenarios. Using these queries, we retrieve raw, untrimmed videos

from popular video-sharing platforms, including YouTube and Bilibili. A metadata filter then uses timestamped comments and danmaku to localize high-dynamic segments within the retrieved videos (Appendix A.5).

Task Definition. We formulate FastBench as a Streaming Video Question Answering (QA) task over an untrimmed, continuous video stream $\mathcal{V}$ arriving online. At a query timestamp t_q, a user question Q is presented, requiring evidence anchored within a temporal window $[t_{\text{start}}, t_{\text{end}}]$. Following the temporal scopes considered in streaming benchmarks (Niu et al., 2025; Lu et al., 2026b), we distinguish three cases: *Backward* ($t_{\text{end}} < t_q$), *Instant* ($t_{\text{start}} \leq t_q \leq t_{\text{end}}$), and *Forward* ($t_{\text{start}} > t_q$). Fig. 2(a) illustrates two of these scopes: the backward example requires recalling whether an F1 tire contacted a track feature before the query, whereas the forward example requires observing and counting a kingfisher's wing flaps after the query.

FastBench explicitly focuses on the unique difficulties of high-dynamic real-world scenarios. Specifically, the benchmark introduces two primary challenges inherent to continuous perception. First, it requires a significantly finer temporal granularity, demanding models to accurately comprehend rapidly evolving visual information across coherent continuous frames rather than isolated snapshots. Second, it highlights a critical trade-off between retaining a long-term historical context and processing high-frequency visual details. Under a limited context token budget, simultaneously accommodating both aspects becomes a severe bottleneck. Given the continuous visual input and query Q, the objective of the model is to effectively navigate this context-granularity trade-off to generate an accurate answer A.

Data Construction Pipeline We designed a comprehensive data construction pipeline that synergizes VLM-driven generation, agentic expert verification, and rigorous human inspection to extract and annotate high-quality question-answering pairs from high-dynamic video streams, as illustrated in Fig. 3. Initially, raw high-resolution videos are segmented into short clips with a duration of no more than 5 seconds. For each segment, we maintain a multi-stage offline conversation with a Vision-Language Model to automate the annotation process. In this initial phase, the VLM analyzes the native high-frame-rate video clip and proposes a candidate high-FPS QA pair.

To ensure the generated questions genuinely require continuous perception of high-dynamic events rather than static scene understanding, we introduce a low-FPS verification mechanism. The video clip is temporally sub-sampled at a sparse rate of 2 FPS and fed back into the VLM along with the candidate question. If the model can correctly answer the question relying solely on this low-FPS clip, the question is identified as pseudo-dynamic and is immediately dropped from the pipeline.

Questions that successfully pass the low-FPS filter then enter the Vision Expert Verification stage. The VLM first performs temporal grounding to isolate the target action interval and selects the initial frame of this segment. It then generates a text query for the target object, which is processed by SAM3 (Carion et al., 2026) to produce a segmentation mask. CoTracker3 (Karaev et al., 2025) tracks the resulting points over the video segment and outputs their trajectory coordinates. The VLM verifies or corrects the original answer from these trajectories. Unverifiable or uncorrectable QA pairs are discarded, while the successful ones advance as double-verified QA pairs.

Human Verification Double-verified QA pairs are mapped back to their longer source videos and reviewed by ten experts in multimodal video understanding. Each sample undergoes three rounds of inspection comparing the original high-FPS video with its 2-FPS subsample to verify: (1) answer correctness; (2) clear, unambiguous visual evidence and question targets; and (3) dependence on high-dynamic visual information, ruling out questions answerable at 2 FPS or without visual context. Experts also annotate the evidence interval and specific timestamps requiring high-FPS perception. Fig. 2(a) illustrates these intervals with Evidence Start and Evidence End markers. Only samples passing all three rounds are retained in FastBench.

3.2 Data Statistics

In this section, we present the overall statistics of the FastBench dataset. We first detail the multi-stage data annotation and filtering process. Subsequently, we analyze the dataset's diversity and comprehensively evaluate its distribution across different capability dimensions and real-world domains. The overall statistical distributions for both capabilities and domains are summarized in Fig. 2(b).

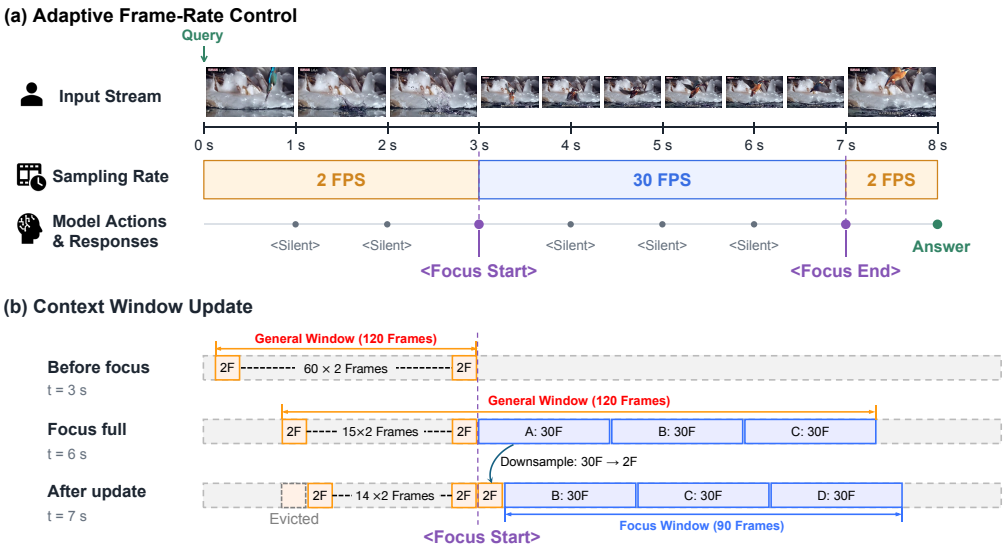


Figure 4: **Overview of ProactiveFrame.** (a) The model switches between low- and high-FPS sampling through `<Focus Start>` and `<Focus End>`; `<Silent>` preserves the current mode. (b) A 120-frame General Window includes a 90-frame Focus Window. When a new high-FPS chunk arrives, the oldest focused chunk is downsampled into sparse history, and the oldest historical chunk is evicted to maintain the budget. Each chunk spans one second; F denotes retained frames.

Data Annotation. To construct FastBench, we employed Gemini-3.1-Pro to generate 89 search keywords for high-dynamic scenarios. For each keyword, a VLM semantically filtered the metadata of the top 200 videos ranked by popularity. This pipeline yielded 766 source videos, segmented into 33,619 clips of up to 5 seconds. The VLM then generated 16,702 candidate question-answering (QA) pairs from these high-frame-rate clips. To ensure temporal dependency and mitigate visual hallucinations, low-FPS filtering and Vision Expert Verification reduced the pool to 3,102 double-verified QAs. Finally, multi-round human inspections for answer correctness, visual clarity, and high-dynamic relevance yielded 306 high-quality samples.

Capability Taxonomy and Distribution. FastBench categorizes its 306 QA pairs into six capability dimensions (Fig. 2(b)): (1) *Action & Physical Interaction* (51.96%) covers fine-grained action recognition, physical contact detection, and rapid event counting; (2) *Predictive & Causal Reasoning* (14.38%) evaluates action anticipation and causal inference; (3) *Motion & Spatiotemporal Tracking* (13.40%) tests spatial reasoning and trajectory tracking under rapid motion; (4) *Entity & Visual Perception* (9.48%) assesses object and text recognition under motion blur; (5) *Temporal & State Dynamics* (8.17%) covers temporal sequencing and sudden state changes; and (6) *Streaming & Online Detection* (2.61%) targets real-time localization of action boundaries. Together, these dimensions assess fine-grained perception and higher-level reasoning under streaming constraints.

Domain Coverage and Diversity. FastBench spans eight domains featuring rapid motion and high visual dynamics: (1) *Sports* (29.08%), including motorsports, combat sports, and skateboarding; (2) *Video Games* (18.63%), featuring fast-paced action and esports gameplay; (3) *Performing Arts* (16.01%), including sleight-of-hand magic, card manipulation, and stage music; (4) *Animals* (12.75%), covering swift wildlife movements and hunting; (5) *Lifestyle & Recreation* (8.50%), encompassing acrobatic stunts, pen spinning, and active hobbies; (6) *Transportation* (8.17%), involving driving, drone operations, and aviation; (7) *Science & Technology* (3.59%), focusing on high-speed physics experiments; and (8) *Food & Cooking* (3.27%), covering rapid culinary preparation. This coverage supports evaluation across varied motion semantics and visual environments.

3.3 Proactive Frame

To resolve the conflict between retaining long-term history and capturing fine-grained temporal details under strict context budgets, we introduce ProactiveFrame, a simple yet effective training-free baseline that enables streaming VLMs to autonomously control their future temporal perception

Table 1: Main evaluation results on **FastBench**. Capabilities are ordered from left to right by sample ratio in descending order. Scores are presented in percentage (%). Values in green or red parentheses indicate gains or drops relative to their respective base models.

Model	Capability Breakdown (by Sample Ratio ↓)						Temporal Scope			Overall
	Action & Phys. Inter. (51.96%)	Predictive & Causal (14.38%)	Motion & Spatiotemp. (13.40%)	Entity & Visual (9.48%)	Temporal & Dynamics (8.17%)	Streaming & Online (2.61%)	Forward (54.90%)	Instant (23.86%)	Backward (21.24%)	
Closed-source Models										
Doubao-Seed-2.0-Lite(ByteDance Seed, 2026)	25.0	51.4	16.6	35.2	30.4	42.5	33.2	24.9	25.2	29.5
GPT-5.6-Sol(OpenAI, 2026)	42.6	56.8	57.6	45.5	43.2	30.0	45.1	51.8	44.9	46.7
Gemini-3.5-Flash(Google DeepMind, 2026b)	43.6	67.3	62.9	44.8	52.0	52.5	52.7	48.5	47.7	50.7
Open-source Models										
Aura (8B)(Lu et al., 2026a)	31.1	32.3	47.8	21.4	28.8	20.0	24.9	41.4	40.3	32.1
Qwen3-VL (8B)(Bai et al., 2025)	30.6	34.1	47.3	32.4	32.0	2.5	33.9	34.5	28.3	32.9
MOSS-VL-Realtime (11B)(Wang et al., 2026)	27.3	34.5	27.8	16.6	31.2	35.0	36.1	22.2	13.2	27.9
JoyAI-VL-Interaction (8B)(Yao et al., 2026)	41.9	48.6	49.8	34.5	43.2	70.0	44.2	43.6	44.3	44.1
Proactive Perception Methods										
VideoChat3 (4B)(Li et al., 2026)	15.7	29.5	18.5	19.3	26.4	67.5	37.6	0.0	0.0	20.7
ProactiveFrame (Qwen3-VL 8B)	36.1 (+5.5)	42.7 (+8.6)	47.3 (0.0)	33.8 (+1.4)	34.4 (+2.4)	40.0 (+37.5)	43.0 (+9.1)	32.3 (-2.2)	32.9 (+4.6)	38.3 (+5.4)
ProactiveFrame (Gemini-3.5-Flash)	47.0 (+3.4)	61.4 (-5.9)	56.1 (-6.8)	56.6 (+11.8)	54.4 (+2.4)	62.5 (+10.0)	53.6 (+0.9)	48.5 (0.0)	52.9 (+5.2)	52.2 (+1.5)

granularity. Instead of passive uniform sampling, the model is guided via text prompts to dynamically switch the frame rate of the incoming stream based on the ongoing streaming context.

Under a fixed context window (e.g., 120 frames), a high frame rate severely compresses the retained history: a 2 FPS stream preserves 60 seconds (sixty 1-second chunks), whereas 30 FPS shrinks the temporal receptive field to merely 4 seconds. To preserve sufficient history while enabling high-dynamic perception, we design a dual-tier context sliding window. As illustrated in Fig. 4, assuming a default 1080p resolution, the General Window holds a constant 120 frames, within which a higher-tier Focus Window buffers high-FPS chunks for fleeting events. Its size is adjustable; for example, 90 frames accommodate three consecutive high-FPS (e.g., native 30 FPS) chunks.

When the accumulated high-FPS frames exceed the Focus Window's capacity, the earliest chunk is evicted following a First-In-First-Out (FIFO) principle. Crucially, rather than being discarded, it undergoes graceful temporal degradation: the chunk is down-sampled to a standard low-FPS chunk (e.g., 2 FPS) and enters the lower-tier General Window. This hierarchical management retains a sparse yet continuous long-term history without sacrificing the ability to proactively perceive transient, high-dynamic events.

4 Experiment

4.1 Experimental Setting

Streaming Evaluation Protocol. We follow the streaming inference framework introduced by PhoStream Lu et al. (2026b), where videos are streamed to the model in 1-second chunks. At each step, the model must decide whether to continue observing or generate a final response. To achieve this, for general (non-streaming) models, we employ explicit text prompts instructing them to output a "Silent" word if they require more temporal context to answer the query. Conversely, for native streaming models, we preserve their inherent interaction paradigms, evaluating them using their default, model-specific wait tokens to ensure fair and optimal performance. The quality of the final open-ended answers is assessed using an LLM-as-a-Judge.

Evaluation Criteria Early and late responses are assigned a score of 0. However, high-dynamic real-world streams present a unique challenge: for tasks like short-term counting, models typically cannot react at the exact millisecond an event ends. Instead, they require a brief subsequent observation period to confirm the transition into a static state and ensure the action has fully concluded. To accommodate this natural reasoning process, we adapt the strict evaluation criteria by introducing a relaxed late-response boundary, extending the valid temporal window to 4 seconds after the event concludes.

Context Window Constraints We define a unified maximum context budget of 248,832,000 pixels, which is equivalent to the information capacity of 120 frames at a 1080p resolution. For our proposed ProactiveFrame, we configure the focus window to 90 frames (at default 1080p resolution) to buffer high-FPS visual chunks, reserving the remaining capacity for at least 30 low-FPS frames to retain a continuous, sparse historical context.

4.2 MAIN RESULTS

Table 1 presents the main evaluation results on FastBench, revealing a noticeable performance gap between closed-source and open-source models in handling high-dynamic streaming scenarios. Among the evaluated models, Gemini-3.5-Flash achieves the highest overall accuracy within the closed-source category at 50.7%. In the open-source category, JoyAI-VL-Interaction (8B) leads the performance with an overall score of 44.1%. Other open-source models demonstrate more modest capabilities on this challenging benchmark, with Qwen3-VL (8B) scoring 32.9%, Aura achieving 32.1%, and MOSS-VL-Realtime (11B) yielding an overall score of 27.9%.

Within the scope of proactive perception methods, VideoChat3 (4B) yields a relatively low overall score of 20.7%. This underperformance can likely be attributed to the model's primary focus on learning to dynamically control the spatial dimension rather than the temporal dimension. Furthermore, its limited 4B parameter capacity inherently restricts its ability to adequately process and reason over highly dynamic, fast-paced visual events.

In contrast, our proposed zero-shot approach, ProactiveFrame, demonstrates substantial effectiveness. Integrating ProactiveFrame with the Qwen3-VL (8B) baseline significantly improves the model's overall performance to 38.3%, representing a notable +5.4% overall increase. This enhancement is driven by massive gains in specific capability breakdowns, highlighted by a remarkable +37.5% surge in the "Streaming & Online" category and an +8.6% increase in "Predictive & Causal" reasoning. Furthermore, combining Gemini-3.5-Flash with our proactive perception method surpasses all other evaluated models, achieving a peak overall score of 52.2%. This configuration also yields impressive sub-category gains, most notably a +11.8% jump in "Entity & Visual" perception and a +10.0% boost in "Streaming & Online" tasks.

However, performance drops are also observed in certain sub-categories—for instance, Gemini-3.5-Flash experiences declines in "Motion & Spatiotemp." (-6.8%) and "Predictive & Causal" (-5.9%), while Qwen3-VL exhibits a minor decrease in the "Instant" temporal scope (-2.2%). This suggests that dynamic temporal resampling can occasionally introduce frame-alignment noise or context redundancy during complex spatio-temporal reasoning. Overall, while these results strongly underscore the immense potential of empowering models with autonomous temporal control, they also reveal that current VLMs' proactive perception mechanisms still have substantial room for improvement in precision and stability.

4.3 ABLATION ANALYSIS

Temporal Context Formulation To examine how streaming models utilize temporal context, Table 2 compares three context formulations: (1) *Full 30 FPS*, processing the stream at its native frame rate; (2) *Mixed FPS (Focus)*, an oracle that provides 30 FPS chunks only within human-annotated focus intervals ($[t_{\text{start}}, t_{\text{end}}]$) and 2 FPS otherwise; and (3) *ProactiveFrame*, where the model triggers focus sampling autonomously from the visual stream. ProactiveFrame improves over the base models in Table 1 (38.3% vs. 32.9% for Qwen3-VL (8B) and 52.2% vs. 50.7% for Gemini-3.5-Flash), but a gap remains to the Mixed FPS oracle (40.7% and 57.5%), indicating that current VLMs still struggle to trigger focus mode at optimal timestamps.

Intriguingly, for Qwen3-VL (8B), *Full 30 FPS* (42.1%) surpasses the oracle *Mixed FPS* (40.7%). This reveals a bottleneck in current open-source VLMs: even with perfectly isolated high-FPS focus segments, their fine-grained dynamic understanding remains limited. Consequently, the visual context lost by downsampling background frames outweighs the benefit of longer historical memory.

Temporal scopes further exhibit distinct patterns. On *Forward* queries, Full 30 FPS exceeds Mixed FPS for both models (56.3% vs. 55.4% on Gemini-3.5-Flash and 42.0% vs. 39.4% on Qwen3-VL). Gemini-3.5-Flash peaks under Full 30 FPS, whereas Qwen3-VL peaks under ProactiveFrame (43.0%). *Backward* queries peak under the oracle Mixed FPS for both models (55.4% and 40.6%). This aligns with task requirements: Forward queries observe events immediately after the query and therefore benefit from dense frames, while Backward queries must retrieve past evidence and are thus sensitive to the memory–resolution trade-off that the oracle balances best.

Input Video Frame Rate To analyze the effect of temporal granularity, we evaluate Qwen3-VL (8B) at input frame rates from 2 to 30 FPS under a unified context budget (Table 3). Performance scales monotonically from the 2 FPS baseline (32.9%) up to 24 FPS, peaking at 44.6% (+11.7%),

Table 2: Ablation study on temporal context formulation using Qwen3-VL (8B) and Gemini-3.5-Flash on FastBench. Mixed FPS (Focus) provides original 30 FPS video chunks during the annotated focus timestamps and 2 FPS video chunks otherwise.

Method	Capability Breakdown (by Sample Ratio ↓)						Temporal Scope			Overall
	Action & Phys. Inter. (51.96%)	Predictive & Causal (14.38%)	Motion & Spatiotemp. (13.40%)	Entity & Visual (9.48%)	Temporal & Dynamics (8.17%)	Streaming & Online (2.61%)	Forward (54.90%)	Instant (23.86%)	Backward (21.24%)	
Qwen3-VL (8B)										
Full 30 FPS	41.0	40.9	57.6	37.2	42.4	7.5	42.0	46.3	37.5	42.1
Proactive Frame	36.1	42.7	47.3	33.8	34.4	40.0	43.0	32.3	32.9	38.3
Mixed FPS (Focus)	41.6	41.4	49.3	35.2	36.8	7.5	39.4	43.8	40.6	40.7
Gemini-3.5-Flash										
Full 30 FPS	49.4	61.4	65.9	60.0	56.8	62.5	56.3	58.1	49.5	55.3
Proactive Frame	47.0	61.4	56.1	56.6	54.4	62.5	53.6	48.5	52.9	52.2
Mixed FPS (Focus)	49.2	65.0	69.8	71.0	60.0	60.0	55.4	64.1	55.4	57.5

Table 3: Ablation study on input video frame rate (FPS) using **Qwen3-VL (8B)** on **FastBench**. Capabilities are ordered from left to right by sample ratio in descending order. Scores are presented in percentage (%). Values in green or red parentheses indicate performance gains or drops relative to the 2 FPS baseline.

FPS	Capability Breakdown (by Sample Ratio ↓)						Temporal Scope			Overall
	Action & Phys. Inter. (51.96%)	Predictive & Causal (14.38%)	Motion & Spatiotemp. (13.40%)	Entity & Visual (9.48%)	Temporal & Dynamics (8.17%)	Streaming & Online (2.61%)	Forward (54.90%)	Instant (23.86%)	Backward (21.24%)	
2 FPS	30.6	34.1	47.3	32.4	32.0	2.5	33.9	34.5	28.3	32.9
4 FPS	37.0 (+6.4)	40.5 (+6.4)	50.2 (+2.9)	35.2 (+2.8)	41.6 (+9.6)	2.5 (0.0)	37.4 (+3.5)	41.1 (+6.6)	38.8 (+10.5)	38.6 (+5.7)
8 FPS	40.3 (+9.7)	38.6 (+4.5)	54.1 (+6.8)	33.8 (+1.4)	43.2 (+11.2)	0.0 (-2.5)	38.1 (+4.2)	44.1 (+9.6)	42.5 (+14.2)	40.5 (+7.6)
16 FPS	43.0 (+12.4)	44.5 (+10.4)	59.0 (+11.7)	40.0 (+7.6)	40.0 (+8.0)	2.5 (0.0)	41.8 (+7.9)	45.8 (+11.3)	46.8 (+18.5)	43.8 (+10.9)
24 FPS	42.9 (+12.3)	48.6 (+14.5)	61.0 (+13.7)	39.3 (+6.9)	41.6 (+9.6)	2.5 (0.0)	43.5 (+9.6)	46.0 (+11.5)	46.2 (+17.9)	44.6 (+11.7)
30 FPS	41.0 (+10.4)	40.9 (+6.8)	57.6 (+10.3)	37.2 (+4.8)	42.4 (+10.4)	7.5 (+5.0)	42.0 (+8.1)	46.3 (+11.8)	37.5 (+9.2)	42.1 (+9.2)

confirming that dense sampling is essential for capturing rapid state transitions and fleeting motion cues.

However, native 30 FPS slightly degrades performance to 42.1%. This saturation reflects a core trade-off in context-constrained streaming VLMs: maximal frame rate severely compresses the historical receptive field and introduces substantial inter-frame redundancy, so 24 FPS offers the best balance between context depth and high-frequency detail.

Capability and scope breakdowns support this observation. Motion-intensive tasks such as *Motion & Spatiotemporal Tracking* (+13.7%) and *Predictive & Causal Reasoning* (+14.5%) benefit most, peaking at 24 FPS. Meanwhile, the drop at 30 FPS is most pronounced for *Backward* queries (46.2% at 24 FPS to 37.5%), since their reliance on historical evidence suffers from the shortened temporal window, whereas *Instant* queries remain strong at 30 FPS (46.3%), depending mainly on immediate frame resolution.

5 Conclusion

We introduced FastBench, a benchmark for evaluating streaming VLMs on high-dynamic real-world streams, built with a trajectory-grounded pipeline that filters pseudo-dynamic questions and verifies answers against SAM3 and CoTracker3 trajectories before multi-round human inspection. We also presented ProactiveFrame, a training-free baseline that adjusts future frame rates via text tokens and retains history through a dual-tier sliding window. Experiments show that high-dynamic perception remains largely unsolved: the best model reaches only 50.7%, denser sampling helps but is bounded by the granularity–history trade-off, and ProactiveFrame improves over sparse sampling yet trails oracle-guided focusing, indicating that deciding *when* to look closer is the central open challenge. We hope FastBench and its annotated focus timestamps foster future research on proactive temporal control in streaming video understanding.

AI Use Statement

In this work, we used Gemini-3.1-Pro (Google DeepMind, 2026a) to construct the FastBench question-answer annotations in Figure 3. It proposed the search keywords used to collect high-

dynamic source videos, and then annotated each short high-frame-rate clip: it proposed a candidate question-answer pair, discarded the pair when the question was answerable from a 2 FPS subsample, and otherwise selected the action interval, queried SAM3 for a segmentation mask, and verified or revised the answer against CoTracker3 trajectories. Pairs that could not be verified or corrected were discarded. Gemini-3.5-Flash (Google DeepMind, 2026b) was used to read timestamped comments and danmaku when localizing candidate segments. Every pair retained in the benchmark was afterwards accepted or rejected by human experts. We also used an LLM-as-a-judge to score open-ended answers, following the same protocol as PhoStream (Lu et al., 2026b). We have not used generative AI tools to develop theoretical models, formulate or prove mathematical claims, propose hypotheses, design the method, or translate text. We reviewed the model-generated annotations against the source videos and the trajectory evidence, and human experts made the final decision for every released sample. We take responsibility for the final content of this work, including the annotations and evaluation judgments produced with the aid of generative AI.

REFERENCES

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. URL https://arxiv.org/abs/2511.21631.

ByteDance Seed. Seed2.0 model card: Towards intelligence frontier for real-world complexity. Model card, February 2026. URL https://seed.bytedance.com/en/seed2.

Nicolas Carion, Laura Gustafson, Yuan-Ting Hu, Shoubhik Debnath, Ronghang Hu, Didac Suris Coll-Vinent, Chaitanya Ryali, Kalyan Vasudev Alwala, Haitham Khedr, Andrew Huang, et al. SAM 3: Segment anything with concepts. In *International Conference on Learning Representations*, volume 2026, pp. 138846–138923, 2026.

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 (CVPR)*, pp. 18407–18418, June 2024.

Rohan Choudhury, Jean-Sebastien Dandurand, Kai Qiu, Kshitij Madhav Bhat, Kartik Sharma, Liza Dahiya, Yizhou Zhao, Souraja Kundu, Chun-Hsien Lin, Kris M. Kitani, and László A. Jeni. FPS-Bench: A benchmark for high frame-rate video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18598–18608, June 2026.

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 *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13448–13459, October 2025.

Google DeepMind. Gemini 3.1 Pro model card, February 2026a. URL https://deepmind.google/models/model-cards/gemini-3-1-pro/.

Google DeepMind. Gemini 3.5 Flash model card, May 2026b. URL https://deepmind.google/models/model-cards/gemini-3-5-flash/.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. MotionBench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8450–8460, June 2025.

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 *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3328–3338. IEEE, 2025.

Nikita Karaev, Yuri Makarov, Jianyuan Wang, Natalia Neverova, Andrea Vedaldi, and Christian Rupprecht. CoTracker3: Simpler and better point tracking by pseudo-labelling real videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 6013–6022, October 2025. URL https://openaccess.thecvf.com/content/ICCV2025/html/Karaev_CoTracker3_Simpler_and_Better_Point_Tracking_by_Pseudo-Labelling_Real_Videos_ICCV_2025_paper.html.

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, 2025a. doi: 10.1007/s11432-024-4321-9. URL https://doi.org/10.1007/s11432-024-4321-9.

Wei Li, Bing Hu, Rui Shao, Leyang Shen, and Liqiang Nie. Lion-fs: Fast & slow video-language thinker as online video assistant. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 3240–3251. IEEE, 2025b.

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. URL https://arxiv.org/abs/2607.14935.

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.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 8731–8772, 2024.

Xudong Lu, Yang Bo, Jinpeng Chen, Shuhan Li, Xintong Guo, Huankang Guan, Fang Liu, Dunyuan Xu, Peiwen Sun, Heyang Sun, Rui Liu, and Hongsheng Li. AURA: Always-on understanding and real-time assistance via video streams. *arXiv preprint arXiv:2604.04184*, 2026a. URL https://arxiv.org/abs/2604.04184.

Xudong Lu, Huankang Guan, Yang Bo, Jinpeng Chen, Xintong Guo, Shuhan Li, Fang Liu, Peiwen Sun, Xueying Lee, Wei Zhang, Xue Yang, Rui Liu, and Hongsheng Li. PhoStream: Benchmarking real-world streaming for omnimodal assistants in mobile scenarios. In *International Conference on Machine Learning*, 2026b. URL https://icml.cc/virtual/2026/poster/62888.

Junbo Niu, Yifei Li, 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? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18902–18913, June 2025.

OpenAI. GPT-5.6 Sol. OpenAI API documentation, July 9 2026. URL https://developers.openai.com/api/docs/models/gpt-5-6-sol.

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

Pengyu Wang, Chenkun Tan, Shaojun Zhou, Qirui Zhou, Yanxin Chen, Xingyang He, Huazheng Zeng, Jijun Cheng, Chenghao Wang, Xiaomeng Qian, et al. MOSS-VL technical report. *arXiv preprint arXiv:2608.15045*, 2026. URL https://arxiv.org/abs/2608.15045.

Ruyi Xu, Guangxuan Xiao, Yukang Chen, Liuning He, Kelly Peng, Yao Lu, and Song Han. StreamingVLM: Real-time understanding for infinite video streams. In C. Vondrick, B. Hariharan, C. Raffel, L. Pinto, D. Yang, and A. Faust (eds.), *International Conference on Learning Representations*, volume 2026, pp. 61463–61475, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/file/6445dd88ebb9a6a3afa0b126ad87fe41-Paper-Conference.pdf.

Dingyu Yao, Junhao Zhou, Chenxu Yang, Chuanyu Qin, Xiangyu Zeng, Yifei Li, Haowen Hou, Zheming Liang, Congcong Wang, Kaiwen Tuo, et al. JoyAI-VL-Interaction: Real-time vision-language interaction intelligence. *arXiv preprint arXiv:2606.14777*, 2026. URL https://arxiv.org/abs/2606.14777.

Yuanhan Zhang, Bo Li, Haotian Liu, Yong Jae Lee, Liangke Gui, Di Fu, Jiashi Feng, Ziwei Liu, and Chunyuan Li. LLaVA-NeXT: A strong zero-shot video understanding model. LLaVA official technical blog, April 2024. URL https://llava-vl.github.io/blog/2024-04-30-llava-next-video/.