X-Stream: Exploring MLLMs as Multiplexers for Multi-Stream Understanding

Peiwen Sun[*1], Xudong Lu[*1], Huadai Liu[*3], Yang Bo[2], Dongming Wu[1], Huankang Guan[2], Minghong Cai[1], Jinpeng Chen[2], Xintong Guo[2], Shuhan Li[2], Fang Liu[2], Rui Liu[2], and Xiangyu Yue[✉1]

[1] MMLab, Chinese University of Hong Kong
[2] Huawei Research [3] Hong Kong University of Science and Technology

Fig. 1: Our *X*-Stream, as the first multi-stream streaming benchmark, encompasses a diverse range of scenarios featuring multi-angle, multi-view, and multi-device capabilities. ⚖ and ⚖ mean balanced and imbalanced streams. and mean the same domain and different domain streams. and mean the real-world and synthesized pairs.

Abstract. While video streaming understanding has made significant strides, real-world applications, such as live sports broadcasting, autonomous driving, and multi-screen collaboration, inherently demand continuous, multi-stream interactions. However, existing benchmarks are confined to single-stream paradigms, leaving a critical gap in evaluating online, cross-stream reasoning. To bridge this, we introduce ***X*-Stream**, the **first benchmark** dedicated to **multi-stream** streaming understanding. Comprising 4,220 rigorously curated QA pairs across 932 videos, *X*-Stream evaluates 11 subtasks across multi-window, multi-view, and multi-device scenarios. Crucially, our dataset is constructed using a novel dual-verification pipeline that prevents over-reliance on a single stream. Furthermore, we pioneer the conceptualization of multi-modal large language models (MLLMs) as **naive multiplexers**, systematically evaluating their performance through the lens of Signal Multiplexing Theory. Our extensive online inference experiments reveal a stark reality: state-of-the-art MLLMs struggle significantly with concurrent streams, achieving only ∼50% score and exhibiting a poor proactive ability. Ultimately, *X*-Stream exposes the **trade-off** of current multiplexing schemes, providing both a practical evaluation protocol and empirical guidance for next-generation multi-stream agents. Code and data are released at homepage.

Keywords: Multi-Stream · Video Understanding · Streaming Understanding

1 Introduction

Propelled by the rapid evolution of Large Language Models (LLMs) like ChatGPT [31], Gemini [14], and Claude [2], AI has successfully transitioned from academic research to everyday application. Following this trajectory, recent models [8, 26] are pushing these boundaries further; by incrementally processing incoming text and frames, they unlock real-time understanding and interactive capabilities for long-form, single-stream streaming videos.

Beyond the challenge of single continuous streams, modern real-world perception increasingly demands multi-stream collaboration. This spans a remarkably wide range of applications, from multi-screen coordination in office environments and orchestrating live feeds in sports broadcasting, to cooperative navigation between mobile maps and smart glasses, and the synchronization of shoulder and wrist cameras on robotic arms. For instance, with over 40 distinct broadcast cameras operating at a World Cup game, *"how to automatically select and broadcast the optimal stream during a live football game?"* The breadth of these scenarios underscores the **immense practical potential** of multi-stream perception. Consequently, developing such capabilities across multiple simultaneous video streams has become a critical imperative for next-generation AI systems.

Previous multi-video datasets typically lack streaming characteristics, as well as long-duration, accurately timestamped multi-stream annotations. During data construction, we identify the over-reliance on a single stream (i.e., single-stream shortcut) as a strong impediment to high-quality data. Then, a novel data protocol and pipeline are used to guarantee the necessity and sufficiency of multi-stream inputs. Finally, as illustrated in Fig. 2(a-b), we introduce X**-Stream** (pronounced "extreme"), the **first Multi-Stream Streaming Understanding benchmark**. X-Stream comprehensively evaluates models through 4,220 carefully curated QA pairs spanning 932 videos and 451 takes from diverse domains, including daily life, gaming, sports, and autonomous driving. Specifically, the benchmark systematically assesses 4 multi-stream core capabilities across 3 progressive dimensions encompassing 11 sub-tasks: ranging from foundational multimodal perception (e.g., visual/audio/temporal grounding and counting), to high-level logical cognition (e.g., spatial/causal reasoning and anomaly detection), and ultimately to complex decision-making (e.g., behavior planning). Crucially, a higher score across all levels demands the continuous integration of multi-stream omni-modality cues.

In telecommunication, the process of combining multiple signals into one signal over a limited shared medium is called "multiplexing". Since MLLMs can only handle one token stream at a time, a multiplexer is naturally essential for integrating multiple video streams into one token stream. Therefore, we conceptualize current MLLMs **as naive multiplexers** processing with a bounded "bandwidth" of token processing capacity. To systematically evaluate how models handle concurrent inputs, we develop three distinct multiplexing strategies based on stream division techniques: Spatial, Temporal, and Semantic Division Multiplexing. Finally, we observe the inherent performance trade-offs dictated by these strategies under varying constraints. We reveal that no single approach

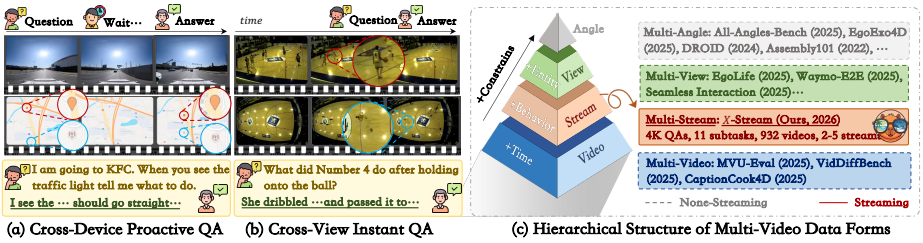


Fig. 2: The illustration of the multi-streaming task. Fig.(a) and (b) showcase the practical examples in daily life. Essentially, the multi-streaming task involves multiple videos with temporal constraints and alignment, requiring the synchronization of video timestamps, as shown in Fig.(c). However, compared to multi-view and multi-angle, it also necessitates important streaming properties to fit the online applications.

is universally optimal; rather, their effectiveness is sensitive to the available token bandwidth and the number of concurrent streams. For instance, while spatial division excels in cross-stream referencing, semantic division becomes more necessary to preserve critical information when scaling to three or more streams under tight token budgets. Finally, within this framework, we conduct comparative evaluations of popular models and ablation studies under online streaming inference conditions.

Overall, this paper yields the following key contributions for multi-stream:

- We propose the **first multi-stream** streaming benchmark, X**-Stream**, including 4,220 carefully curated diverse QAs spanning 932 videos and 451 takes. Most top-performing models achieve only about 50% score, and advanced cross-stream skills, like causal reasoning, remain far from application.
- To address the over-reliance on a single stream during data construction and evaluation, we introduce a novel data protocol and pipeline that guarantees the **necessity and sufficiency** of multi-stream inputs. Accordingly, our X-Stream benchmark heavily prioritizes the model's multi-stream capabilities.
- We also systematically observe the **inherent trade-offs** introduced by different strategies in multi-stream video multiplexing. Furthermore, we provide a comprehensive analysis to guide future architectural designs.

2 Related Works

2.1 Multimodal Large Language Models

MLLMs have garnered significant attention, driving the emergence of exceptional application-level products. In the realm of video understanding, closed-source models such as GPT-5 [31], Gemini 3 Pro [14], and Doubao-2.0 [7] currently achieve state-of-the-art performance. Concurrently, open-source models, including InternVL 3.5 [36], MiniCPM-V 4.5 [55], Qwen 3.5 [27], and DeepSeek-VL2 [45], have demonstrated highly competitive capabilities. This rapid progress spans various sub-domains of video analysis, ranging from general comprehension [13,56] to spatial [19,32,41] and temporal reasoning [10,11]. Nevertheless, despite these remarkable perceptual breakthroughs, a critical limitation persists: most existing MLLMs are confined to the offline processing of multiple complete videos, lacking the capability to perform online inference on continuous multiple video streams.

2.2 Streaming Understanding

Streaming requires models to perceive real-time interactions, track forward audio-visual inputs, and respond in the right time. Pioneering efforts [12, 23, 43] in the community initially focused on audio streaming capabilities, successfully achieving application-level interaction.

However, streaming video understanding has emerged more recently, primarily constrained by the challenges of processing extensive video tokens. Consequently, advancements are categorized into modeling architectures and data resources. On the modeling front, research focuses on efficiency and interaction. Frameworks like VideoLLM-online [8], StreamingVLM [51], and Streamo [46] optimize memory and processing efficiency for streaming dialogues. Meanwhile, Dispider [26] addresses perception-reaction conflicts through an asynchronous architecture, and MMDuet2 [38] employs multi-turn reinforcement learning to enable autonomous decisions on whether to respond or remain silent. On the data front, efforts are divided between training resources and evaluation benchmarks. Large-scale training datasets such as HoloAssist [37] and EgoBlind [47] facilitate the development of streaming understanding capabilities. For evaluation, StreamingBench [20], OVO-Bench [22], PhoStream [21], OmniMMI [40], and SVBench [53] assess general streaming understanding and proactive reasoning, while ProactiveVideoQA [39] specifically targets user experience in proactive interaction scenarios. Despite these advancements, existing benchmarks predominantly focus on single stream understanding, leaving a notable gap in specialized evaluations for multi-stream streaming scenarios with open-ended model interactions.

2.3 Multi-Video & Multi-View Understanding

From the perspective of video forms, we conceptualize the multi-video data family as a pyramid, illustrated in Fig. 2, organized by increasing constraints: 1) Multi-Video, 2) Multi-Stream, 3) Multi-View, and 4) Multi-Angle.

At the base, **Multi-Video** Understanding, including MVU-Bench [25] and video-differencing [5, 44] represents the task with the fewest constraints, encompassing multi-video understanding across virtually all scenarios. Further narrowing the scope, **Multi-View** Understanding, like EgoLife [52], Wod-e2e [50], Seamless-interaction [1], NuPrompt [42] and NuPlanQA [24] mandates multiple perspectives of the same activity—which may involve different subjects—such as front and rear views in autonomous driving or distinct participant feeds in video chats. Finally, at the peak, **Multi-Angle** Understanding, represented by Assembly101 [30], EgoExo4D [15], and All-Angle Bench [54], imposes the strictest constraints by necessitating different angles of the same subject at the same time, exemplified by front versus top-down views during assembly tasks or synchronized first-person and third-person perspectives.

In the middle of the hierarchy, **Multi-Stream** introduces the critical constraint of timestamp alignment. Although Multi-Angle and Multi-View constitute merely a small fraction of the broader Multi-Stream video category in the pyramid, prior approaches [16, 34] have focused on understanding based on **entire** multi-view video files rather than evaluating in an **online streaming** or even real-

time manner. Consequently, the field of multi-stream streaming understanding remains unexplored.

3 Data Construction

In this section, we present a comprehensive construction protocol and statistical analysis of our X-Stream benchmark, including data collection in Sec. 3.1, task definition in Sec. 3.2, annotation pipeline in Sec. 3.3, and statistical analysis in Sec. 3.4. Further details are available in the appendix.

3.1 Data Collection and Sources

To construct the X-Stream benchmark, we systematically gather data across multi-angle, multi-view, and multi-device configurations, as illustrated in Fig. 1. Our collection comprises 857 hours of raw multi-stream data, featuring 2 to 10 concurrent video streams drawn from over 20 sources. As illustrated in Fig. 4, these sources span eight major domains: driving, sports, robotics, daily routine, chat, surveillance, live streaming, and interface. This raw collection relies on three primary strategies: 1) reformatting metadata from well-established datasets (e.g., Egolife [52], Seamless Interaction [1]); 2) combining existing data with simulation techniques to generate multi-device scenarios (e.g., Comma2K-19 [29] with a dashboard); and 3) manually collecting and recording public source data (e.g., Split-screen Game, Map-Street). After preprocessing and pairing, we select 160 hours of diverse data across 2-5 streams for further processing. Due to page limit, further source details and visual previews are available in the appendix.

3.2 Task Definition

Multi-stream streaming understanding demands that a model perceive queries in time, continuously track multiple information streams, and deliver responses at precisely the right moment. Within this dynamic framework, queries are fundamentally categorized by their temporal requirements into instant and forward questions [20, 21, 39]. **Instant questions** allow the model to generate an immediate response by leveraging retrospective or current context. In contrast, **forward questions** function as proactive tasks where the necessary conditions for an answer have not yet occurred. For these, the model must actively monitor the incoming streams and wait until the appropriate criteria are met before responding. To better evaluate the capabilities of multi-stream models, we systematically model this framework from the dual perspectives of fundamental skills and progressive tasks below.

From a skills perspective, this framework encompasses **4 core capabilities** essential for navigating complex multi-stream environments, as shown in Fig. 3. Single-stream understanding, as the foundational ability, involves accurately extracting precise information from one specific data stream while operating within a broader multi-stream context. Building upon this, the framework requires cross-stream anti-interference to maintain accuracy by actively filtering out contradictory or irrelevant noise from concurrent streams. Furthermore, it

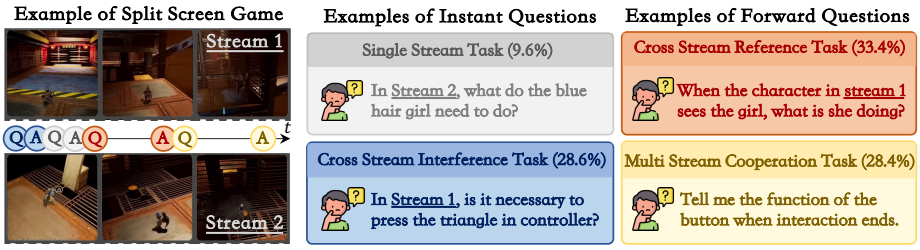


Fig. 3: The illustration of the 4 multi-stream abilities. To evaluate these abilities, our X-Streaming Benchmark includes 3 progressive dimensions and 11 subtasks.

necessitates cross-stream reference alignment to accurately map abstract references in one stream to their corresponding concrete entities or timestamps in another. Finally, the framework culminates in cross-stream cooperation, which demands synthesizing fragmented clues distributed across multiple streams to deduce answers that no single stream could provide alone.

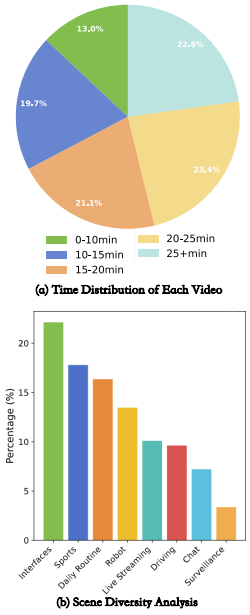
To evaluate these capabilities, the X-Stream benchmark is systematically categorized into **11 progressive tasks**. The foundational level focuses on multimodal perception, encompassing five specific task types: visual, audio, and temporal grounding, counting, as well as saliency detection. As the complexity increases, the benchmark evaluates high-level logical cognition through five distinct tasks: 3D spatial, causal, counterfactual, and commonsense reasoning, alongside anomaly detection. At the highest level of complexity, the benchmark tests decision-making capabilities, specifically focusing on behavior planning. Crucially, across all three dimensions, generating accurate answers strictly requires the continuous integration of multi-stream cues.

3.3 Data Pipeline

To generate high-quality QA pairs, we follow the pipeline outlined in Alg. 1, which consists of four main stages: preprocessing, QA generation, sufficiency and necessity verification, and human verification.

Preprocessing. To establish a baseline video stream for all subsequent generation steps, we first resample the video to 2 FPS. Compared to Gemini's default 1 FPS, this higher sampling density minimizes the loss of fast-moving actions. Next, we divide and resize the 2 FPS MP4 file into segments smaller than 50MB. This chunking process ensures reliable parallel processing while strictly adhering to storage and transmission limits.

QA generation with time stamps. We employ a hybrid approach: automated generation via MLLMs with rejection sampling, alongside template-based generation for specific subtasks. To construct knowledge- and action-intensive tasks, we randomly sample videos from our dataset and leverage ground-truth metadata, where available, to draft initial questions based on rich curated templates. Finally, we use the Gemini-3-Pro model to refine these template-based questions, rendering them more natural and challenging, and give an accurate answer with a rationale and distractors (if needed). Throughout the construction



Algorithm 1: X-streams Benchmark Pipeline

Input: RawVideo **Output:** Multi-Stream QA Benchmark

```
MultiStreams = Preprocess(RawVideo);                          } Multi-Stream Pre-process
AllCandQA = EmptySet;
FinalQA = EmptySet;
for Video in MultiStreams do                                  } QA Generation
    CandQA = GenerateQA(Video);
    Append(AllCandQA, CandQA);
end
for QA in AllCandQA do                                        } Multi-Stream Sufficiency
    Clip = TrimVideo(MultiStreams, QA.Timestamp);
    if Check(Clip, QA.Question) == Correct then
        Retain = True;
        for SingleStream in Clip do                           } Multi-Stream Necessity
            Ans = Check(SingleStream, QA.Question);
            if Ans == Correct then
                Retain = False;
                break;
            end
        end
        if Retain == True then
            Append(FinalQA, QA);
        end
    end
end
FinalQA = HumanCheck(FinalQA);                                } Human Modification
return FinalQA;
```

Fig. 4: Diversity analysis.

phase, a balanced type and task distribution is ensured. Additionally, we maintain a roughly 1:1 ratio between multiple-choice and free-form QAs.

Shortcut observation. During the verification process, we observe that the model tends to generate "pseudo multi-stream" QA pairs that inherently rely on single-stream information. While factually correct, these QAs fail to genuinely evaluate the model's cross-stream capabilities. This degradation primarily manifests in two forms: **1) Pseudo reference**, where cross-stream anchoring becomes invalid. For example, suppose an ongoing action in Stream 1 (e.g., 'sitting') spans the entire video. In this case, querying it during a momentary event in Stream 2 becomes trivial, making temporal alignment meaningless. **2) Pseudo cooperation**, which arises from high information redundancy across streams (e.g., overlapping fields of view). In such cases, the model can resolve the query using any single stream, eliminating the need for genuine collaboration or information complementarity.

QA verification. To mitigate the hallucination and the above shortcut, we implement a dual-verification process based on Sufficiency and Necessity. **Multi-stream Sufficiency** addresses error pairs by ensuring a powerful model can correctly answer the question when given the video clipped to the verified timestamp. Conversely, **Multi-stream Necessity** eliminates shortcuts by verifying that the model completely fails to answer if provided with only isolated, individual streams. During verification, we utilize the videos of high resolution and bit rate, since the target timestamp is already fixed. The final results must satisfy both the necessity and sufficiency criteria. Together, this approach guarantees high answer accuracy while strictly requiring joint multi-stream understanding.

Human verification and modification. To ensure data quality and safety, we recruit 31 video understanding experts to conduct a rigorous two-round review. During this process, experts verify the clarity of the questions and the accuracy and completeness of the answers, ensuring that responses relied exclusively on audio-visual evidence. Specifically, instant questions had to be fully supported by content preceding the timestamp, whereas forward-looking questions required the timestamp to mark the earliest possible moment the question became answerable, without any premature information leakage. When identifying flawed samples, experts resolve these issues by editing the QA text, correcting task labels, or adjusting timestamps. Conversely, we discard samples that were ambiguous, open to multiple interpretations, insufficiently supported by evidence, or lacking expert consensus. Ultimately, following this meticulous two-stage validation and filtering process, we retain only high-quality samples that were accurate, clear, safe, and strictly grounded. Additionally, the workers' interface, human verification statistics, and error type distribution are provided in the appendix.

3.4 Benchmark Statistics

As the pioneering multi-stream streaming benchmark, X-Stream is specifically designed to handle complex multi-stream interactions and real-world application scenarios. The benchmark features an accurate dataset comprising 4,220 QAs, 932 videos, and 451 takes. To better fit actual streaming environments, video durations are kept between 5 and 30 minutes, with an average length of 15.8 minutes. While dual-stream videos form the core of the benchmark, approximately 20% of the data consists of takes with 3 to 5 streams to support more comprehensive multi-stream analysis. Furthermore, around 30% of the questions incorporate audio or speech information. As shown in Tab. 1, compared to other benchmarks, ours demonstrates significant advantages in multi-stream and streaming capabilities. Meanwhile, other statistics remain comparable to other popular datasets in QA and video tasks. Together, these comprehensive metrics demonstrate that X-Stream is well-equipped for evaluating advanced, multi-modal streaming applications.

Table 1: Statistics of multi-domain video/data sources used in our study. "∼" means "approximately". Note: one "take" consists of multiple videos, while an individual video may be reused across multiple "takes".

Tasks	Dataset	QA	Videos	Videos Per Take	Duration (h)	Cross Stream	Open Ended	Streaming	Proactive
Multi-View & Multi-Video	EgoLife-Eval [52]	0.3K	1	6	20	✗	✗	✗	✗
	ProMQA-Assembly [16]	0.4K	0.2K	2	7	✗	✓	✗	✗
	WaymoQA [50]	6.4K	∼1K	2	∼2	✓	✓	✗	✗
	MVU-Bench [25]	1.8K	5K	3-5	15	✗	✗	✗	✗
	VidDiff [5]	4.5K	0.5K	2	3	✗	✓	✗	✗
Streaming	OVO-Bench [22]	2.8K	0.6K	1	85	✗	✗	✓	✓
	StreamingBench [20]	4.5K	0.9K	1	136	✗	✗	✓	✓
	Inf-Streams-Eval [51]	2.5K	0.5K	1	42	✗	✓	✓	✗
	LiveSports [9]	1.2K	0.8K	1	40	✗	✓	✓	✓
	ProactiveVideoQA [39]	1.4K	1.4K	1	49	✗	✓	✓	✓
	OmniMMI [40]	2.3K	1.1K	1	100	✗	✓	✓	✓
	MMDuet [38]	2.0K	2.0K	1	100	✗	✓	✓	✓
	ESTP-Bench [58]	2.3K	1.2K	1	80	✗	✓	✓	✓
	PhoStream [21]	5.6K	0.6K	1	92	✗	✓	✓	✓
Multi-Stream	X-Stream (Ours)	4.2K	0.9K	2-5	160	✓	✓	✓	✓

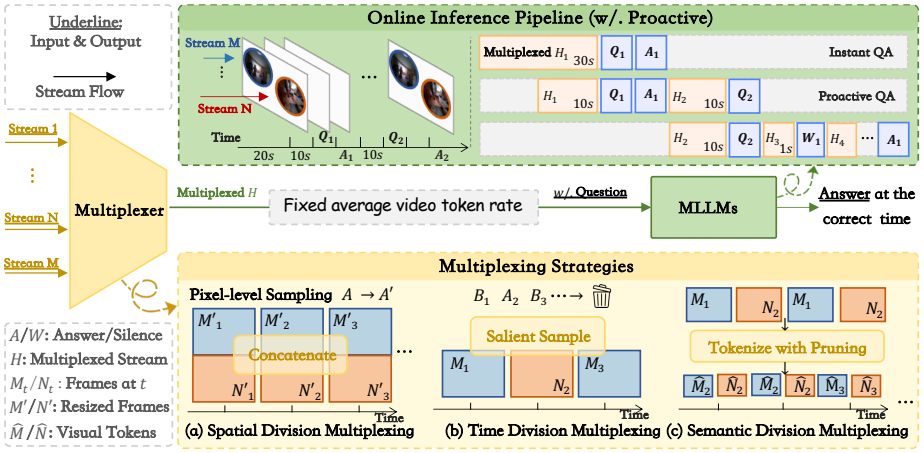


Fig. 5: MLLMs can only handle one token stream at a time, making a multiplexer essential for integrating multiple video streams into one token stream. To address this, we investigate three multiplexing strategies and uncover their inherent trade-offs. During evaluation, the model sequentially processes continuous video streams in 1-second intervals while maintaining a sliding memory window for context management.

More Information on our X-Stream Benchmark. We encourage readers to consult the appendix for further information, including but not limited to comprehensive source investigations, self-developed data-collection tools, in-depth distribution analyses, rigorous quality control, licensing terms, and QA previews.

4 MLLMs as Naive Multiplexers

> ***Preliminary 1:*** Multiplexing [17] is a fundamental method in telecommunication by which multiple signals are combined into one signal over a limited shared medium. The method prevents signal interference by dividing them by orthogonality, such as time, frequency, and wavelength.

Since MLLMs can only take one **token stream** at a time, integrating multiplexing into multiple **video streams** allows us to systematically analyze the approaches to processing multiple streams. In practical multi-stream scenarios, MLLMs are inherently constrained by limited context windows and computational budgets. To reflect these real-world limitations, similar to channel bandwidth in telecommunications, a **limited and fixed average video token rate**, denoted as C_{max}, is always enforced.

We investigate three multiplexing strategies and analyze their inherent trade-offs. The demonstration below illustrates this process using a two-stream setup for clarity, with two concurrent frames M_t and N_t at time t.

1) Spatial Division Multiplexing in Fig. 5(a). By leveraging the spatial separability of pixels within the frame, this method directly stitches two video streams together and feeds them into MLLMs as a single stream. Formally, we apply a spatial downsampling function $D(\cdot, r)$ with retention ratios r_m and r_n. The combined input is constructed by pixel level concatenation as $X_t =$

Concat$(D(M_t, r_m), D(N_t, r_n))$, subject to the system's token capacity constraint $|\mathcal{T}(X_t)| \leq C_{max}$, where $\mathcal{T}(\cdot)$ denotes the tokenization process. However, this approach requires video re-encoding and audio overlapping prior to input, which introduces extra processing. We also provide grid layout analysis in the appendix.

2) Time Division Multiplexing in Fig. 5(b). By processing different video streams as independent inputs, concurrent frames across these streams are assigned identical temporal embedding. Unlike orthogonality mentioned in the preliminary, MLLMs rely on a specific stream identifier (`<stream N>`) to achieve this kind of separation. Mathematically, we introduce binary indicator variables $\alpha_t, \beta_t \in \{0, 1\}$ to determine whether a frame from stream M or N is sampled at time t. This sampling is constrained by the token budget: $\alpha_t|\mathcal{T}(M_t)|+\beta_t|\mathcal{T}(N_t)| \leq C_{max}$. To ensure temporal consistency between M_t and N_t, we explicitly align their temporal embeddings. In practice, however, achieving this synchronized temporal encoding is only feasible with open-source models.

3) Semantic Division Multiplexing in Fig. 5(c). Unlike the previous methods that operate on physical dimensions (space and time), this approach multiplexes streams within the semantic space. Mathematically, the basic idea is to formulate a semantic selection function $\mathcal{S}(\mathcal{T}(\cdot), k)$ that retains the k most salient tokens from a given stream, with the constraint $|\mathcal{S}(\mathcal{T}(M_t), k_m)| + |\mathcal{S}(\mathcal{T}(N_t), k_n)| \leq C_{max}$ where k_m and k_n are the allocated token quotas for streams M and N. To implement $\mathcal{S}$, following previous training-free token pruning methods [33, 57], we optimally balance token similarity and diversity to maintain salient information with low latency. With the help of extra visual encoders [28, 35], a conditional Determinantal Point Process (DPP) kernel matrix is constructed for the candidate tokens as:

$$K_{ij} = \text{relevance}_i \cdot \text{similarity}_{ij} \cdot \text{relevance}_j. \tag{1}$$

where "relevance" denotes how relevant the token is to the current query, and "similarity" represents the degree of similarity between the two tokens. Then, a greedy Maximum A Posteriori (MAP) inference algorithm is employed to iteratively select the subsets of size k_m and k_n. In each stream, the algorithm selects the token with the highest marginal gain and dynamically penalizes the scores of remaining candidates that share high similarity with the selected one. This rigorous selection mechanism ensures that the retained tokens are both highly relevant to the task and visually diverse. Finally, the tokens from different streams are interleaved via time division above. Since we cannot evaluate proprietary models by the token-level modification, as a workaround, we convert the stream with the most retained tokens back into frames to use as input.

Leveraging these multiplexing schemes, multiple video streams are integrated into a unified token sequence before being input into VLLM for inference.

5 Experiments

In this section, we present a series of experiments on X-Stream. We describe the experimental setup, report baseline results, conduct a multiplexing ablation study, and perform a human test to support the LLM-as-a-Judge evaluation.

5.1 Experiment Setup

Following [21], we evaluate three categories of baseline models with the Online Inference Pipeline and LLM-as-a-Judge evaluation. We report Instant, Backward, Forward, and comprehensive scores for all models. For further analysis of the Forward setting, we also report the proportions of Early Response (ER, ↓) and No Response (NR, ↓). ER denotes any response other than Silent or a placeholder that occurs before Timestamp Proactive. NR denotes that the model produces no response other than Silent or a placeholder within the response window. Then, we use a 2-second response window and run inference for 6 time slots. Therefore, achieving a high score on a forward question requires providing the correct answer at the precise moment. However, when calculating the score of multi-stream abilities, we only average the scores of temporally accurate answers to avoid imbalance caused by different answer timings. Additionally, we cap the average $C_{max} = 250$ tokens per video second. However, due to variations in token calculation across models, we employ diverse methods to enforce this limit, such as adjusting playback speed and resizing videos. r_m, r_n are dynamically set at the largest value within C_{max}. Also, α_t, β_t are uniformly sampled from $\{0, 1\}$.

Table 2: Comprehensive streaming performance comparison of mLLMs on the X-Stream Benchmark (Ours). The symbols "▶" and "🔊" indicate video and audio support, respectively. Background colors denote the top three results within each scene: green (1st), blue (2nd), and yellow (3rd). Among open-source models, **bold** and underlined highlight the best and second-best.

Model	Overall	Evaluation Score (↑)				Forward Time		Multi-Stream Abilities			
		Instant	Backward	Forward	Compre.	ER (↓)	NR (↓)	Single Stream	Multi Coop.	Cross Ref.	Cross Inter.
Human Preference ▶🔊	91.84	91.73	95.19	85.10	97.50	9.50	2.55	94.12	92.05	90.10	98.55
Proprietary Multimodal Models											
Gemini 3 Pro [14] ▶🔊	49.60	73.38	72.23	20.77	82.04	73.13	0.23	72.45	71.16	74.79	66.96
GPT-5 [31] ▶	27.78	44.28	37.18	6.51	59.83	81.73	1.14	39.08	44.12	52.75	45.65
GPT-4o [18] ▶	22.46	37.28	32.72	4.05	47.01	87.14	0.74	34.83	34.90	43.77	37.52
Doubao-Seed-1.8 [6] ▶	36.79	55.49	57.18	14.52	59.13	66.19	3.95	47.55	35.69	56.52	60.82
Open-source Multimodal Models											
Qwen2.5-VL-7B [4] ▶	25.49	40.02	36.02	8.34	45.28	68.10	11.36	43.80	41.43	42.72	40.01
Qwen2.5-Omni-7B [48] ▶🔊	26.82	41.96	41.17	9.03	45.04	53.19	22.51	38.60	40.80	41.86	44.35
Qwen3-VL-8B [3] ▶	26.78	43.41	33.30	7.53	51.01	78.40	6.50	49.88	43.41	33.30	51.01
Qwen3-Omni-30B-A3B [49] ▶🔊	**34.28**	**63.92**	**53.40**	0.61	**69.16**	98.81	**0.27**	**63.41**	55.68	**66.08**	56.58
Qwen3-VL-30B-A3B [3] ▶	34.19	52.09	38.54	**14.46**	57.26	73.91	1.18	54.68	**57.90**	65.98	**65.22**
Open-source Streaming Models											
Dispider [26] ▶	15.44	21.71	19.29	8.09	23.90	55.63	7.26	38.01	21.97	23.65	31.37
VideoLLM-online-8B [8] ▶	8.48	15.00	15.53	0.03	17.67	99.10	0.66	13.15	16.90	10.15	6.70
MMDuet2 [38] ▶	6.79	11.76	10.37	1.44	11.27	**31.49**	54.11	15.84	14.44	9.16	4.96

5.2 Main Results

Performance comparison of streaming abilities. In Tab.2, we adopt the straightforward Spatial Division Multiplexing for our primary comparative experiments. As shown in Tab.2, proprietary models consistently outperform their open-source counterparts across all settings on the full *X*-Stream benchmark. Notably, while Qwen3-Omni-30B-A3B leads the open-source models due to its strong comprehension, its overall Forward capability is hindered by suboptimal response timing. Furthermore, open-source streaming models generally underperform, constrained by limited training data and difficulties in handling frequent

Table 3: Comprehensive performance comparison of MLLMs across 3 multi-stream dimensions and 11 core tasks on X-Stream Benchmark. The "$*$" symbol indicates that the model lacks audio capabilities and is evaluated directly on the question, while other notations follow Tab. 2.

Model	Foundational Grounding					Logical Cognition					Agency
	Visual Grd.	Audio Grd.	Temporal Grd.	Object Count.	Saliency Detect.	3D spa.	Counter-factual	Causal Reasoning	Common Sense	Anomaly Detect.	Decision-Making
Proprietary Multimodal Models											
Gemini 3 Pro [14]	66.72	64.82	68.93	76.37	63.61	70.82	75.00	41.79	69.35	70.52	44.18
GPT-5 [31]	36.68	21.63*	36.64	42.52	53.55	52.28	15.00	37.65	44.77	45.54	28.74
GPT-4o [18]	31.99	22.15*	32.83	36.19	42.14	40.74	40.00	33.33	38.04	37.86	24.53
Doubao-Seed-1.8 [6]	49.87	29.91	49.31	52.75	61.13	59.14	85.00	52.45	57.92	54.29	37.10
Open-source Multimodal Models											
Qwen2.5-VL-7B	38.90	27.82*	40.12	30.16	44.26	45.22	40.00	36.27	40.02	33.04	21.19
Qwen3-VL-8B [3]	46.03	25.84*	47.26	46.60	49.82	48.95	20.00	38.24	42.65	35.89	30.18
Qwen3-Omni-30B-A3B [49]	53.61	**52.15**	**64.56**	**64.77**	**68.61**	**63.29**	**90.00**	**53.14**	**64.08**	**60.18**	27.14
Qwen3-VL-30B-A3B [3]	**64.33**	29.83*	54.55	54.68	60.96	60.45	40.00	40.22	50.84	41.25	**31.59**
Open-source Streaming Models											
Dispider [26]	19.58	13.80*	18.67	25.10	22.50	22.12	50.00	17.20	21.65	17.32	10.94
VideoLLM-online-8B [8]	12.62	17.97*	21.47	12.90	22.64	21.94	0.00	18.14	21.53	21.13	14.44
MMDuet2 [38]	22.27	18.35*	22.24	36.10	24.14	23.63	10.00	17.90	22.16	20.42	9.12

Table 4: The impact of multiplexing ablation on individual abilities on all X-Stream Benchmark.

Model	Scheme	Single Stream	Multi-Coop	Cross-Ref	Cross-Inter
Qwen3-Omni-30B-A3B	Spatial	63.41	55.68	**66.08**	56.58
	Time	**69.78**	**58.76**	58.62	**69.14**
	Semantic	61.58	52.13	55.34	59.03
Gemini-3-Pro	Spatial	72.45	71.16	**74.79**	66.96
	Time	**79.62**	**75.13**	67.08	**80.74**
	Semantic	70.30	66.64	70.92	68.94

Table 5: Performance comparison of the number of streams under different multiplexing schemes.

Model	Scheme	N=2	N=3	N=4	N=5
Qwen3-Omni-30B-A3B	Spatial	36.61	36.88	34.22	25.84
	Time	**40.55**	36.17	35.83	25.44
	Semantic	36.15	**38.46**	**40.64**	**29.82**
Gemini-3-Pro	Spatial	57.47	19.86	30.48	22.81
	Time	**58.09**	24.19	31.76	22.14
	Semantic	55.50	**26.89**	**41.25**	**31.06**

proactive queries. Consequently, we select the leading proprietary model, Gemini 3 Pro, alongside top-performing open-source models for the in-depth observational experiments detailed below.

Performance comparison of multi-stream abilities. In Fig. 3, X-Stream's three-dimensional evaluation reveals a clear multi-stream capability hierarchy across current MLLMs, transitioning from basic perception to complex reasoning. When calculating dimension subtask and ability scores, timing effects are excluded to prevent outliers from affecting the results, therefore, yielding a higher overall score. Models generally exhibit robust performance in foundation tasks, but struggle significantly with advanced cognitive demands. Specifically, decision-making and specific logical cognition tasks like causal reasoning emerge as the most formidable bottlenecks, yielding lower scores across all model tiers. Overall, X-Stream perfectly serves as the multi-stream evaluation standard.

> ***Preliminary 2:*** In signal theory, each scheme presents inherent limitations. 1) Frequency division → waste of resources; 2) time division → latency and jitter; and 3) code division → interference between codes.

Analysis of multiplexing scheme. In a hypothetical scenario with a large number of streams, spatial or time division would degrade into full blurriness or discontinuity, losing all usable information; semantic division, however, would still manage to retain a basic semantic content. Conversely, in a single-stream

Table 6: Performance comparison of audio settings. In spatial division, we simply superimpose the audio tracks.

Modality	Qwen3-Omni -30B-A3B		Gemini-3 -Pro	
	Spatial	Time	Spatial	Time
w/. audio	**34.28**	26.57	**49.60**	41.52
w/o. audio	29.40	**30.37**	46.13	**45.84**

Table 7: Performance comparison in multi-stream setting. Videos from single-stream datasets are combined with distractor videos to form pseudo multi-stream.

Model	Input	Multi-Stream	Single Stream		
		X-Stream (Ours)	Streaming- -Bench	OVO -Bench	Pho- -Stream
Qwen3-Omni -30B-A3B	Single Stream	5.35	**57.19**	**56.16**	**33.01**
	Multiple Stream	**34.28**	26.10	34.78	17.33
Gemini 3 Pro	Single Stream	13.93	**65.20**	**58.91**	**44.64**
	Multiple Stream	**49.60**	35.91	38.05	13.96

scenario, the outcome of spatial and time division becomes standard streaming video, whereas semantic division would introduce unnecessary semantic loss. Based on our extended evaluations, we summarize the distinct advantages of each multiplexing approach:

1) Spatial Division Multiplexing excels in temporal modeling and cross-stream referencing. As shown in Tab.4, it demonstrates superior cross-stream referencing in multi-stream setups. This likely occurs because representing multiple frames within a single unit preserves the model's pretrained inference dynamics and temporal perception.

2) Time Division Multiplexing is optimal under relaxed token rate constraints. As shown in Tab. 4 and 7, evaluations reveal that this method thrives in dual-stream scenarios, which form the main part of our X-Stream Bench. Furthermore, by circumventing the severe selection penalties of Semantic Division, it better preserves visual details and delivers stronger performance in 2-stream scenarios.

3) Semantic Division Multiplexing dominates under strict token constraints and high stream counts (≥ 3 streams). As Tab.3 illustrates, scaling up the number of streams causes Spatial and Time Division representations to become severely blurred or fragmented. Under these dense information loads, Semantic Division effectively filters and preserves critical semantic features.

These empirical findings are **similar to** the time-tested theory of signal multiplexing. Our exploration of multi-stream video multiplexing can serve as the foundation of future multi-streaming understanding models.

Analysis of multiplexing ablation on audio. Tab. 4 shows that audio-capable methods perform significantly better in audio grounding, highlighting the importance of audio. However, Tab.6 reveals that the impact of multiplexing varies significantly between audio and video. For instance, Spatial Division Multiplexing often causes multi-channel audio to overlap. Conversely, while Time-Division resolves this overlap, it introduces semantic discontinuity in speech. Since audio signals possess stronger inherent coupling than image pixels, our discussion is limited to simple multiplexing techniques.

Necessity of multi-stream. Tab.7 presents further results validating the necessity of multi-stream processing. Single-stream inference fails on our multi-stream benchmark. Conversely, injecting distracting streams into single-stream datasets causes severe performance degradation. This confirms that our X-Stream benchmark evaluates a fundamentally novel capability, rather than merely extending single-stream systems.

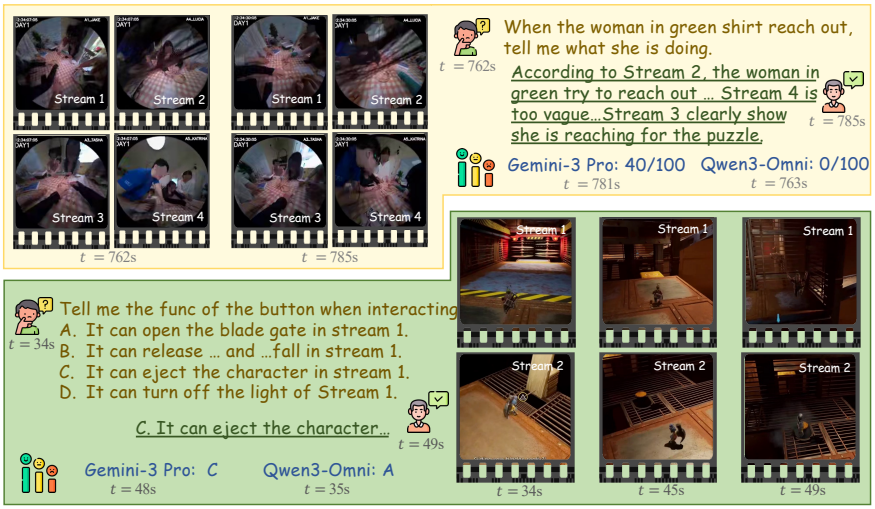


Fig. 6: The case study in our X-Stream Benchmark. We choose a 4-stream, proactive, free-form QA (yellow) and a 2-stream, proactive, multi-choice QA (green) as examples.

Human verification of LLM-as-judge: To validate the effectiveness, we request both the LLM-as-judge and human experts to evaluate 200 QAs. The Spearman correlation of 0.62 ($p<0.05$) confirms that LLM-as-judge reliably mirrors human evaluation. See the appendix for prompt and model details.

More experiments and analysis. As shown in Fig.6, our X-Stream benchmark requires the integration of multi-stream information to generate timely and accurate answers. We encourage readers to consult the appendix for more details, including case preview, experiments, and analysis.

6 Discussion and Conclusion

Discussion: Despite its wide range of applications, the full potential of multi-stream remains untapped. On the data front, public video datasets are not precise enough without film-standard professional gear and expert work, usually causing streams to drift out of sync over time. On the technical side, current multiplexing strategies still struggle to balance video comprehension with temporal reasoning.
Conclusion: In this paper, we introduce X-Stream, the first comprehensive benchmark dedicated to multi-stream streaming understanding. To cover as many real-world scenarios as possible, we develop a rigorous dual-verification pipeline, ensuring that our 4,220 curated QA pairs genuinely demand cross-stream understanding. By evaluating popular MLLMs as naive multiplexers, our extensive experiments reveal that current models still struggle with continuous multi-stream integration, achieving only around 50% accuracy and falling short in proactive tasks. Furthermore, we systematically analyze the inherent trade-offs of three multiplexing strategies under varying token constraints and stream counts. Ultimately, X-Stream exposes the limitations of existing streaming architectures, providing both a robust evaluation framework and empirical guidance for designing the next generation of efficient, real-time multi-stream agents.

Acknowledgments

This work is partially supported by HK RGC-Early Career Scheme (No. 24211525), and ITSP Platform Project (No. ITS/600/24FP). This study was supported in part by the Centre for Perceptual and Interactive Intelligence, a CUHK-led InnoCentre under the InnoHK initiative of the Innovation and Technology Commission of the Hong Kong Special Administrative Region Government. This work is also partially supported by Hong Kong RGC Strategic Topics Grant (No. STG1/E-403/24-N), and a CUHK-CUHK(SZ)-GDST Joint Collaboration Fund (No. YSP26-4760949). This work is also partially supported by the Huawei Hong Kong Research Center (HKRC).

References

1. Agrawal, V., Akinyemi, A., Alvero, K., Behrooz, M., Buffalini, J., Carlucci, F.M., Chen, J., Chen, J., Chen, Z., Cheng, S., et al.: Seamless interaction: Dyadic audio-visual motion modeling and large-scale dataset. arXiv preprint arXiv:2506.22554 (2025) 4, 5
2. Anthropic: Claude sonnet 4.5 system card. System card, Anthropic (Sep 2025), https://www-cdn.anthropic.com/963373e433e489a87a10c823c52a0a013e9172dd.pdf 2
3. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., Ge, W., Guo, Z., Huang, Q., Huang, J., Huang, F., Hui, B., Jiang, S., Li, Z., Li, M., Li, M., Li, K., Lin, Z., Lin, J., Liu, X., Liu, J., Liu, C., Liu, Y., Liu, D., Liu, S., Lu, D., Luo, R., Lv, C., Men, R., Meng, L., Ren, X., Ren, X., Song, S., Sun, Y., Tang, J., Tu, J., Wan, J., Wang, P., Wang, P., Wang, Q., Wang, Y., Xie, T., Xu, Y., Xu, H., Xu, J., Yang, Z., Yang, M., Yang, J., Yang, A., Yu, B., Zhang, F., Zhang, H., Zhang, X., Zheng, B., Zhong, H., Zhou, J., Zhou, F., Zhou, J., Zhu, Y., Zhu, K.: Qwen3-vl technical report (2025), https://arxiv.org/abs/2511.21631 11, 12
4. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., Zhong, H., Zhu, Y., Yang, M., Li, Z., Wan, J., Wang, P., Ding, W., Fu, Z., Xu, Y., Ye, J., Zhang, X., Xie, T., Cheng, Z., Zhang, H., Yang, Z., Xu, H., Lin, J.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025) 11
5. Burgess, J., Wang, X., Zhang, Y., Rau, A., Lozano, A., Dunlap, L., Darrell, T., Yeung-Levy, S.: Video action differencing. arXiv preprint arXiv:2503.07860 (2025) 4, 8
6. ByteDance: Doubao-seed-1.8. Available at Volcengine ARK Platform (2026) 11, 12
7. ByteDance Seed Team: Seed2.0 model card: Towards intelligence frontier for real-world complexity. https://lf3-static.bytednsdoc.com/obj/eden-cn/lapzild-tss/ljhwZthlaukjlkulzlp/seed2/0214/Seed2.0 Model Card.pdf (2026), accessed: 2026-02-26 3
8. Chen, J., Lv, Z., Wu, S., Lin, K.Q., Song, C., Gao, D., Liu, J.W., Gao, Z., Mao, D., Shou, M.Z.: 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) 2, 4, 11, 12
9. Chen, J., Zeng, Z., Lin, Y., Li, W., Ma, Z., Shou, M.Z.: Livecc: Learning video llm with streaming speech transcription at scale. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 29083–29095 (2025) 8

10. Chen, J.J., Liao, Y.C., Lin, H.C., Yu, Y.C., Chen, Y.C., Wang, F.: Rextime: A benchmark suite for reasoning-across-time in videos. Advances in Neural Information Processing Systems **37**, 28662–28673 (2024) 3
11. Cheng, Z., Hu, J., Liu, Z., Si, C., Li, W., Gong, S.: V-star: Benchmarking video-llms on video spatio-temporal reasoning. arXiv preprint arXiv:2503.11495 (2025) 3
12. Défossez, A., Mazaré, L., Orsini, M., Royer, A., Pérez, P., Jégou, H., Grave, E., Zeghidour, N.: Moshi: a speech-text foundation model for real-time dialogue. arXiv preprint arXiv:2410.00037 (2024) 4
13. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., 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) 3
14. Google DeepMind: Gemini 3 pro model card. Available at Google DeepMind Model Cards (2025) 2, 3, 11, 12
15. Grauman, K., Westbury, A., Torresani, L., Kitani, K., Malik, J., Afouras, T., Ashutosh, K., Baiyya, V., Bansal, S., Boote, B., et al.: Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 19383–19400 (2024) 4
16. Hasegawa, K., Imrattanatrai, W., Asada, M., Holm, S., Wang, Y., Zhou, V., Fukuda, K., Mitamura, T.: Promqa-assembly: Multimodal procedural qa dataset on assembly. arXiv preprint arXiv:2509.02949 (2025) 4, 8
17. Haykin, S.: Communication Systems. John Wiley & Sons, Inc., New York, 4th edn. (2001) 9
18. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024) 11, 12
19. Lang, S., Liu, J., He, H., Sun, P., Chen, Y., Liu, T., Yang, L., Guo, L., Zhang, H.: Longspace: Exploring long-horizon spatial memory from perception to recall in video. arXiv preprint arXiv:2606.05677 (2026) 3
20. Lin, J., Fang, Z., Chen, C., Wan, Z., Luo, F., Li, P., Liu, Y., Sun, M.: Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. arXiv preprint arXiv:2411.03628 (2024) 4, 5, 8
21. Lu, X., Guan, H., Bo, Y., Chen, J., Guo, X., Li, S., Liu, F., Sun, P., Li, X., Zhang, W., et al.: Phostream: Benchmarking real-world streaming for omnimodal assistants in mobile scenarios. arXiv preprint arXiv:2601.22575 (2026) 4, 5, 8, 11
22. Niu, J., Li, Y., Miao, Z., Ge, C., Zhou, Y., He, Q., Dong, X., Duan, H., Ding, S., Qian, R., et al.: Ovo-bench: How far is your video-llms from real-world online video understanding? In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18902–18913 (2025) 4, 8
23. OpenAI: Gpt realtime, https://developers.openai.com/api/docs/models/gpt-realtime 4
24. Park, S.Y., Cui, C., Ma, Y., Moradipari, A., Gupta, R., Han, K., Wang, Z.: Nuplanqa: A large-scale dataset and benchmark for multi-view driving scene understanding in multi-modal large language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 8066–8076 (2025) 4
25. Peng, T., Wang, H., Zhang, Y., Wang, Z., Wang, Z., Chang, G., Yang, J., Li, S., Wang, Y., Wang, X., et al.: Mvu-eval: Towards multi-video understanding evaluation for multimodal llms. arXiv preprint arXiv:2511.07250 (2025) 4, 8

26. Qian, R., Ding, S., Dong, X., Zhang, P., Zang, Y., Cao, Y., Lin, D., Wang, J.: Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24045–24055 (2025) 2, 4, 11, 12
27. Qwen Team: Qwen3.5: Towards native multimodal agents (February 2026), https://qwen.ai/blog?id=qwen3.5 3
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021) 10
29. Schafer, H., Santana, E., Haden, A., Biasini, R.: A commute in data: The comma2k19 dataset. arXiv preprint arXiv:1812.05752 (2018) 5
30. Sener, F., Chatterjee, D., Shelepov, D., He, K., Singhania, D., Wang, R., Yao, A.: Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21096–21106 (2022) 4
31. Singh, A., Fry, A., Perelman, A., Tart, A., Ganesh, A., El-Kishky, A., McLaughlin, A., Low, A., Ostrow, A., Ananthram, A., et al.: Openai gpt-5 system card. arXiv preprint arXiv:2601.03267 (2025) 2, 3, 11, 12
32. Sun, P., Lang, S., Wu, D., Ding, Y., Feng, K., Liu, H., Ye, Z., Liu, R., Liu, Y.H., Wang, J., et al.: Spacevista: All-scale visual spatial reasoning from mm to km. arXiv preprint arXiv:2510.09606 (2025) 3
33. Tang, C., Ek, S., Koch, D., Mullins, R.D., Weddell, A.S., Chauhan, J.: Surge: Surprise-guided token reduction for efficient video understanding with vlms. In: The Fourteenth International Conference on Learning Representations 10
34. Tian, S., Wang, R., Guo, H., Wu, P., Dong, Y., Wang, X., Yang, J., Zhang, H., Zhu, H., Liu, Z.: Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. arXiv preprint arXiv:2506.13654 (2025) 4
35. Tschannen, M., Gritsenko, A., Wang, X., Naeem, M.F., Alabdulmohsin, I., Parthasarathy, N., Evans, T., Beyer, L., Xia, Y., Mustafa, B., et al.: Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. arXiv preprint arXiv:2502.14786 (2025) 10
36. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025) 3
37. Wang, X., Kwon, T., Rad, M., Pan, B., Chakraborty, I., Andrist, S., Bohus, D., Feniello, A., Tekin, B., Frujeri, F.V., et al.: Holoassist: an egocentric human interaction dataset for interactive ai assistants in the real world. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 20270–20281 (2023) 4
38. Wang, Y., Liu, S., Wang, D., Xu, N., Wan, G., Zhang, H., Zhao, D.: Mmduet2: Enhancing proactive interaction of video mllms with multi-turn reinforcement learning. arXiv preprint arXiv:2512.06810 (2025) 4, 8, 11, 12
39. Wang, Y., Meng, X., Wang, Y., Zhang, H., Zhao, D.: Proactivevideoqa: A comprehensive benchmark evaluating proactive interactions in video large language models. arXiv preprint arXiv:2507.09313 (2025) 4, 5, 8
40. Wang, Y., Wang, Y., Chen, B., Wu, T., Zhao, D., Zheng, Z.: Omnimmi: A comprehensive multi-modal interaction benchmark in streaming video contexts. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 18925–18935 (2025) 4, 8

41. Wu, D., Liu, F., Hung, Y.H., Duan, Y.: Spatial-mllm: Boosting mllm capabilities in visual-based spatial intelligence. arXiv preprint arXiv:2505.23747 (2025) 3
42. Wu, D., Han, W., Liu, Y., Wang, T., Xu, C.z., Zhang, X., Shen, J.: Language prompt for autonomous driving. In: Proceedings of the AAAI conference on artificial intelligence. vol. 39, pp. 8359–8367 (2025) 4
43. Wu, D., Wang, T., Zhang, Y., Zhang, X., Shen, J.: Onlinerefer: A simple online baseline for referring video object segmentation. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2761–2770 (2023) 4
44. Wu, J., Li, S., Bian, Z., Chen, J., Wen, R., Ping, A., He, Y., Wang, J., Zhang, Y., Liu, J.: Vidic: Video difference captioning. arXiv preprint arXiv:2512.03405 (2025) 4
45. Wu, Z., Chen, X., Pan, Z., Liu, X., Liu, W., Dai, D., Gao, H., Ma, Y., Wu, C., Wang, B., et al.: Deepseek-vl2: Mixture-of-experts vision-language models for advanced multimodal understanding. arXiv preprint arXiv:2412.10302 (2024) 3
46. Xia, J., Chen, P., Zhang, M., Sun, X., Zhou, K.: Streaming video instruction tuning. arXiv preprint arXiv:2512.21334 (2025) 4
47. Xiao, J., Huang, N., Qiu, H., Tao, Z., Yang, X., Hong, R., Wang, M., Yao, A.: Egoblind: Towards egocentric visual assistance for the blind people. arXiv preprint arXiv:2503.08221 (2025) 4
48. Xu, J., Guo, Z., He, J., Hu, H., He, T., Bai, S., Chen, K., Wang, J., Fan, Y., Dang, K., Zhang, B., Wang, X., Chu, Y., Lin, J.: Qwen2.5-omni technical report (2025), https://arxiv.org/abs/2503.20215 11
49. Xu, J., Guo, Z., Hu, H., Chu, Y., Wang, X., He, J., Wang, Y., Shi, X., He, T., Zhu, X., Lv, Y., Wang, Y., Guo, D., Wang, H., Ma, L., Zhang, P., Zhang, X., Hao, H., Guo, Z., Yang, B., Zhang, B., Ma, Z., Wei, X., Bai, S., Chen, K., Liu, X., Wang, P., Yang, M., Liu, D., Ren, X., Zheng, B., Men, R., Zhou, F., Yu, B., Yang, J., Yu, L., Zhou, J., Lin, J.: Qwen3-omni technical report (2025), https://arxiv.org/abs/2509.17765 11, 12
50. Xu, R., Lin, H., Jeon, W., Feng, H., Zou, Y., Sun, L., Gorman, J., Tolstaya, E., Tang, S., White, B., et al.: Wod-e2e: Waymo open dataset for end-to-end driving in challenging long-tail scenarios. arXiv preprint arXiv:2510.26125 (2025) 4, 8
51. Xu, R., Xiao, G., Chen, Y., He, L., Peng, K., Lu, Y., Han, S.: Streamingvlm: Real-time understanding for infinite video streams. arXiv preprint arXiv:2510.09608 (2025) 4, 8
52. Yang, J., Liu, S., Guo, H., Dong, Y., Zhang, X., Zhang, S., Wang, P., Zhou, Z., Xie, B., Wang, Z., et al.: Egolife: Towards egocentric life assistant. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 28885–28900 (2025) 4, 5, 8
53. Yang, Z., Hu, Y., Du, Z., Xue, D., Qian, S., Wu, J., Yang, F., Dong, W., Xu, C.: Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. arXiv preprint arXiv:2502.10810 (2025) 4
54. Yeh, C.H., Wang, C., Tong, S., Cheng, T.Y., Wang, R., Chu, T., Zhai, Y., Chen, Y., Gao, S., Ma, Y.: Seeing from another perspective: Evaluating multi-view understanding in mllms. arXiv preprint arXiv:2504.15280 (2025) 4
55. Yu, T., Wang, Z., Wang, C., Huang, F., Ma, W., He, Z., Cai, T., Chen, W., Huang, Y., Zhao, Y., et al.: Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. arXiv preprint arXiv:2509.18154 (2025) 3
56. Yu, Z., Xu, D., Yu, J., Yu, T., Zhao, Z., Zhuang, Y., Tao, D.: Activitynet-qa: A dataset for understanding complex web videos via question answering. In: Proceedings of the AAAI Conference on Artificial Intelligence. pp. 9127–9134 (2019) 3

57. Zhang, Q., Liu, M., Li, L., Lu, M., Zhang, Y., Pan, J., She, Q., Zhang, S.: Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. arXiv preprint arXiv:2506.10967 (2025) 10
58. Zhang, Y., Shi, C., Wang, Y., Yang, S.: Eyes wide open: Ego proactive video-llm for streaming video. arXiv preprint arXiv:2510.14560 (2025) 8