StreamOV: Streaming Omni-Video Understanding via Evidence-Guided Memory and Response Triggering

Ming Xie[1,2*] **Zizheng Huang**[1,3*] **Xudong Tan**[2*] **Chao Wang**[2*]
Xiangyu Zeng[3] **Wenxiao Wu**[1,5] **Tao Chen**[1,2†] **Limin Wang**[3,4†] **Yanwei Fu**[1,2†]
[1]Shanghai Innovation Institute [2]Fudan University [3]Nanjing University
[4]Shanghai Artificial Intelligence Laboratory [5]Huazhong University of Science and Technology

Abstract

While streaming omni-video understanding demands continuous perception and proactive, real-time interaction, this crucial area remains largely under-explored. Current omni-modal methods are inherently designed for offline settings, limiting their applicability in streaming scenarios due to two fundamental flaws. First, they lack robust mechanisms to manage continuously growing audio-visual context over long horizons and cannot autonomously initiate responses at opportune moments. Second, existing benchmarks are predominantly confined to offline, single-turn question answering, failing to capture continuous, multi-turn streaming interactions. To bridge these gaps, we propose StreamOV, a novel **Stream**ing **O**mni-**V**ideo understanding framework for efficient online audio-visual reasoning with bounded memory and proactive response triggering. Specifically, StreamOV introduces a multimodal evidence-guided long-short term memory that condenses historical audio-visual context into compact informative evidence under a fixed budget. It further employs a hidden-state-driven trigger to decide when to respond, avoiding explicit silence-token generation and external routers. We also curate SOVBench, the first comprehensive benchmark for online, multi-turn omni-modal evaluation. Extensive experiments show that StreamOV achieves state-of-the-art performance across diverse streaming and omni-video benchmarks, demonstrating its effectiveness for both online and offline video understanding.

1 Introduction

Video understanding is moving from offline perception toward continuous interaction. In real-world scenarios such as long-form multimedia streams, a model cannot assume access to a complete video before answering. Instead, it must process synchronized visual and audio signals as they arrive, maintain useful historical context under bounded computation, and determine whether the current evidence is sufficient to generate a response. We refer to this setting as *streaming omni-video understanding*, where the model is expected to perform online audio-visual perception, long-horizon reasoning, and proactive interaction over a continuous multimodal stream.

Despite rapid progress in large video and omni-modal models [1, 2, 3, 4], most existing systems remain designed for offline inference, where a fixed set of frames, audio segments, or the full video context is provided before generation. Such a formulation is fundamentally mismatched with streaming scenarios: the audio-visual context grows continuously, making naive context accumulation inefficient, and the model lacks an explicit mechanism for deciding *when* to respond. Existing streaming video understanding methods partially address online perception through buffers, recurrent

*Equal contribution: mxie24@m.fudan.edu.cn
†Corresponding author: yanweifu@fudan.edu.cn, lmwang@nju.edu.cn

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

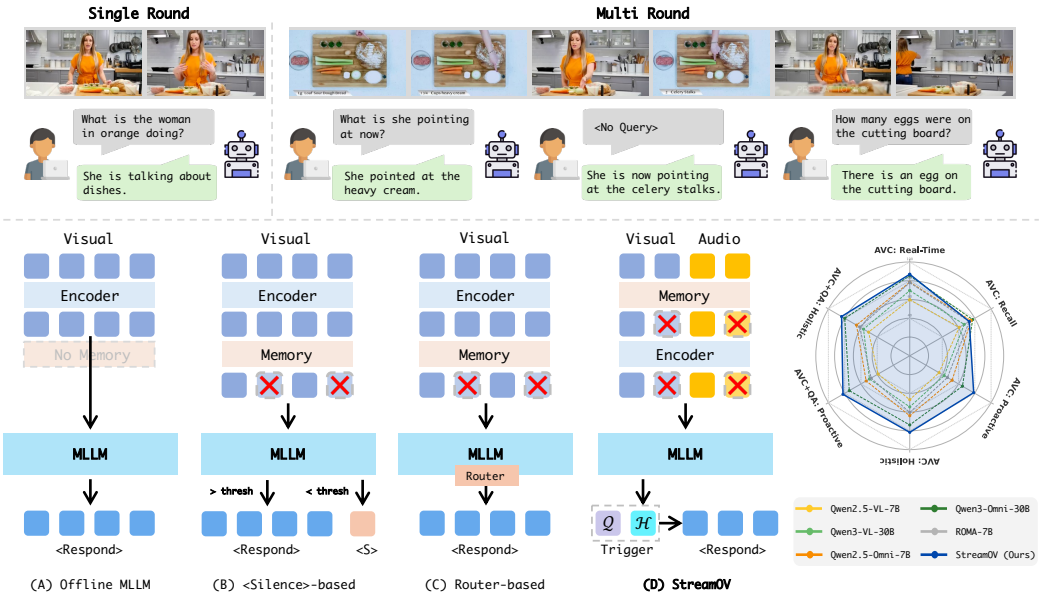


Figure 1: **Overview of StreamOV and SOVBench.** Streaming omni-video understanding requires historical multimodal memory and proactive response decisions. Compared with offline MLLMs, silence-based methods, and router-based methods, StreamOV uses compact multimodal memory and a lightweight trigger to respond efficiently without an external router. The radar chart shows state-of-the-art performance across multiple dimensions.

states, or token compression [5, 6, 7, 8, 9], but they are predominantly visual-centric and overlook synchronized acoustic evidence. As shown in Fig. 1, existing proactive streaming methods usually rely on silence-token generation [5, 10, 11] or external routers [12, 13]. Silence tokens introduce non-semantic control outputs that may perturb the pretrained generation distribution, while routers delegate response decisions to auxiliary models with limited multimodal reasoning.

Evaluation also remains insufficient. Existing video benchmarks mainly focus on offline or single-turn question answering, while current streaming benchmarks either emphasize visual-only perception or evaluate audio-visual QA as independent single-round tasks. They therefore fail to measure whether a model can preserve long-range multimodal context, handle multi-round temporal interactions, infer latent user intent from dialogue history, and intentionally remain silent when queried evidence is absent. To bridge these gaps, we propose **StreamOV**, a streaming omni-video understanding framework built on an omni-modal large language model. StreamOV constructs multimodal evidence from both query-agnostic stream dynamics and query-aware semantic relevance, routes it into visual-only, audio-only, and audio-visual-aligned cues, and maintains a long-short term memory that preserves dense recent observations while sparsely retaining informative historical evidence under a fixed budget. For proactive interaction, StreamOV further introduces an MLLM-as-a-trigger mechanism that probes the hidden states of early decoding steps to predict whether to *Respond* or *Wait*, avoiding both explicit `<silence>` token generation and external routers.

We further construct **SOVBench**, the first benchmark dedicated to online multi-round omni-video evaluation. SOVBench-O evaluates continuous audio-visual comprehension across Real-Time, Recall, and Proactive interaction paradigms, while SOVBench-T directly evaluates operational proactivity as a response-triggering task, requiring models to respond when queried evidence appears and remain silent when it is absent. Extensive experiments on SOVBench, StreamingBench, OVO-Bench, and offline audio-visual benchmarks demonstrate that StreamOV achieves strong performance across streaming omni-modal, visual-only, and offline settings.

Our contributions are summarized as follows:

- We formulate *streaming omni-video understanding* as an online audio-visual interaction problem requiring bounded memory, long-horizon reasoning, and proactive response timing, and con-

struct **SOVBench**, the first benchmark for online multi-round omni-video evaluation with both comprehension and triggering tasks.

- We propose **StreamOV**, which performs multimodal evidence routing and long-short term memory to efficiently preserve informative visual, audio, and audio-visual cues under streaming setting.
- We introduce a hidden-state-driven response trigger that leverages the MLLM's early decoding states to decide whether to respond or wait, avoiding explicit silence-token generation and external routers while achieving state-of-the-art or competitive performance across multiple benchmarks.

2 Related Works

Offline Video Understanding. Large Vision-Language Models (LVLMs) have driven rapid progress in offline video understanding. Early works (e.g., VideoChat [14], Video-LLaVA [15], LLaVA-NeXT-Video [16]) adapted image-based LVLMs via sparse frame sampling and spatial-temporal modeling. This visual-centric approach recently culminated in models like Qwen-VL series [17, 18, 3, 4], which excel at fine-grained spatial-temporal reasoning over extended context windows. However, recognizing video's multi-sensory nature, the field is now shifting toward omni-modal understanding. Moving beyond simplistic late-fusion approaches, unified end-to-end architectures like Qwen-Omni series [1, 2] natively interleave audio, vision, and text. This enables joint cross-modal reasoning to accurately capture the intricate correlations between visual actions and acoustic events.

Streaming Omni-Video Understanding. Compared with offline video reasoning and understanding, streaming video understanding requires low latency and efficient memory management [10, 19]. Early works [5, 6] established this paradigm with streaming buffers, while subsequent studies improve long-context modeling through KV cache compression and recurrent memory [7, 8]. Computational efficiency has been further optimized via adaptive resource allocation [20] and hierarchical token processing [21, 9, 22]. For interactive scenarios, recent models [23, 24, 25] enable real-time dialogue and continuous perception. However, most existing streaming video understanding methods still focus primarily on the visual modality, leaving synchronized acoustic streams underexplored [26]. In this work, we study Streaming Omni-Video Understanding, where models must continuously perceive, reason, and interact over synchronized visual and audio streams under an online setting.

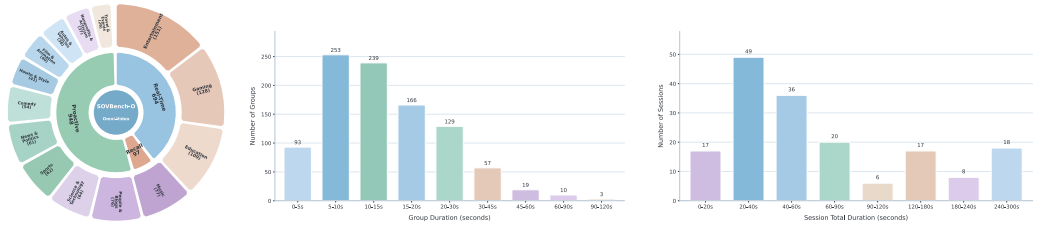
Online Video Benchmarks. MovieChat-1K [27], VideoLLM-online [5], OVBench [6] and ODVBench [9] pioneered long-term memory and streaming evaluation protocols, while OVOBench [28] and VStream-QA [21] introduced objective metrics to ensure causal consistency in real-time perception. More recently, StreamingBench [29] and SVBench [30, 31] established comprehensive benchmarks for streaming reasoning and multi-round interactions. However, while StreamingBench incorporates audio-visual QA, it is limited to independent single-round tasks; SVBench, on the other hand, supports multi-round interactions but lacks comprehensive audio-visual grounding. Consequently, there remains a critical gap in evaluating a model's multi-round online audio-visual QA capabilities.

3 SOVBench

To comprehensively evaluate streaming omni-video understanding, we introduce SOVBench, which consists of two complementary components. SOVBench-O evaluates continuous comprehension accuracy through multi-round temporal interactions, while SOVBench-T assesses the model's proactive response decision-making and intentional silence capabilities.

3.1 Data Source Filtering & Generation

To construct a robust benchmark for streaming omni-video understanding (SOVU), we select FineVideo [32] as our primary data source. FineVideo provides large-scale internet videos paired with exceptionally dense metadata, including global descriptions, scene-level narratives, localized activities, prop interactions, and timestamped Automatic Speech Recognition transcripts. These fine-grained, multimodal annotations serve as an ideal foundation for evaluating continuous comprehension in a streaming context. While FineVideo offers extensive annotations, the complexity of SOVU necessitates data with dense information, and strong cross-modal alignment. To ensure the highest quality of our benchmark and eliminate noisy or trivial samples, we employ an advanced large



(A) Question Category (B) Group-Level Duration (C) Session-Level Duration

Figure 2: Statistics of SOVBench-O. (a) Distribution of question categories across different real-world domains. (b) Distribution of group-level (QAs with follow-up reasoning) durations, where each group corresponds to a temporally coherent local interaction segment. (c) Distribution of session-level durations, reflecting the temporal span of complete streaming sessions.

Table 1: Comparison with existing video understanding benchmarks. SOVBench is designed for streaming omni-video understanding, covering audio-visual perception, multi-turn interaction, proactive response, and intentional silence.

Benchmark	Online Streaming	Multi-turn Interaction	Audio Input	ProactiveResponse	Silence Decision	Omni-video Understanding
Video-MME	✗	✗	✓	✗	✗	✗
MLVU	✗	✗	✗	✗	✗	✗
LongVideoBench	✗	✗	✗	✗	✗	✗
EgoSchema	✗	✗	✗	✗	✗	✗
StreamingBench	✓	✗	✓	✗	✓	✓
OVO-Bench	✓	✗	✗	✗	✓	✗
SVBench	✓	✓	✗	✗	✗	✗
SOVBench (ours)	✓	✓	✓	✓	✓	✓

language model [4] as an automated quality assessor to filter the dataset based on its raw metadata. We design a rigorous evaluation prompt that scores each video on a scale of 0 to 10 across five critical dimensions: Visual Dynamism, Narrative Coherence, Information Density, Audio-Visual Alignment and Reasoning Value (see Supp. B for details). Alongside the quantitative scores, the assessor generates detailed rationales. Finally, we apply a strict score threshold for filtering and conduct manual sanity checks on these rationales to guarantee the reliability of the selection process. To subsequently generate high-quality multi-round QAs, we extract the fine-grained multimodal metadata from the filtered videos and reorganize it into a strictly chronological timeline integrating visual events, acoustic signals, and transcripts. We then prompt Gemini [33] with this temporal blueprint to formulate multi-round interactions that are strictly grounded in the timeline while covering short-term occurrences, long-term historical dependencies, and cross-modal reasoning. The generated QAs are manually verified by annotators to ensure temporal correctness and cross-modal grounding.

3.2 SOVBench-O: Multi-round Omni QA Benchmark

To comprehensively evaluate audio-visual capabilities, we construct SOVBench-O, a benchmark containing multi-round QAs targeting fine-grained visual details (*e.g.*, scenes, appearances, expressions, text) and key auditory elements (*e.g.*, acoustic events, speech). While some questions assess single-modality capabilities, a subset of questions requires joint multimodal reasoning. To strictly evaluate continuous streaming comprehension, we design three temporal interaction paradigms: **Real-Time**, **Recall**, and **Proactive**. Real-Time QA tests on-the-fly understanding of immediate, ongoing events. Recall QA tests long-term memory retention by querying information that appeared earlier in the video stream. Finally, Proactive QA assesses autonomous contextual awareness across two distinct dimensions: *Cognitive Proactivity* and *Operational Proactivity*. Within SOVBench-O, we specifically evaluate cognitive proactivity by testing implicit context understanding. In such scenarios, explicit questions are entirely omitted and only candidate options are provided if the current turn is highly correlated with the previous round, forcing the model to infer the latent intent of the user from short-term dialogue history. Through these diverse paradigms, SOVBench-O provides a rigorous evaluation of cognitive comprehension and contextual reasoning in continuous streams.

3.3 SOVBench-T: Omni Response Triggering Benchmark

Beyond content accuracy, the streaming nature of SOVU demands exceptional operational proactivity (optimal triggering judgment), since knowing whether to respond is as critical as what to say. Existing approaches often incorporate specialized tokens, such as `<silence>` [10, 5], into the training process to endow the model with the capacity for proactive response and intentional silence. However, their evaluations typically rely on auxiliary tasks [12, 34], such as temporal grounding, to indirectly reflect the effectiveness of these mechanisms. To provide a more intuitive evaluation of a model's proactive response capability in a video stream, we construct SOVBench-T, a benchmark centered on a binary classification task to evaluate this fundamental ability of online models. Specifically, positive samples consist of content where the queried omni-information eventually appears, requiring the model to autonomously trigger a timely response within a temporal window. In contrast, negative samples involve queries for information that is entirely absent, forcing the model to demonstrate its capacity for intentional silence and refrain from generating hallucinated responses. By focusing on this binary verification paradigm, SOVBench-T provides a direct measure of whether a model's proactive behavior is grounded in actual omni-perceptions.

3.4 Benchmark Statistics

As shown in Fig. 2, after filtering and manual verification, SOVBench-O contains 172 streaming sessions from 172 distinct FineVideo videos, comprising 1,739 question-answer turns organized into 969 temporally coherent dialogue groups. Each group corresponds to a short local interaction segment, while successive groups within the same session evaluate the model's ability to preserve long-range context over time. These groups cover both single-turn and multi-turn interactions, with many requiring follow-up reasoning rather than isolated one-shot recognition. On average, each group contains 1.79 turns and spans 15.46 seconds of video content, with a median duration of 12.08 seconds. Among all groups, 368 are single-turn interactions, while 601 contain multiple turns, indicating that a substantial portion of the benchmark evaluates follow-up reasoning. Beyond temporal diversity, SOVBench-O spans 15 top-level categories and 86 fine-grained semantic categories across diverse real-world domains, preventing the evaluation from collapsing into narrow action recognition and requiring models to reason over varied audio-visual evidence, narration styles, and conversational intents. SOVBench-T further focuses on response triggering in real-world streaming scenarios. It contains 226 valid samples, with 120 positive samples requiring a response and 106 negative samples where the model should remain silent, forming a near-balanced binary evaluation set. More statistics are provided in Supp. A.

3.5 Benchmark Analysis

Tab. 1 highlights the key differences between SOVBench and existing video understanding benchmarks. *Silence Decision* is query-conditioned: given an existing question, the model must determine whether the currently accumulated audio-visual evidence is sufficient to answer, and remain silent when the evidence is still insufficient or never appears. *Proactive Response* is query-free: when a new stream segment arrives without a new question, the model must infer the user's continuing intent from the preceding interaction and autonomously initiate an updated response when appropriate. Most offline benchmarks evaluate video comprehension with fixed input videos and predefined questions, making them insufficient for measuring online interaction ability. Recent streaming benchmarks move toward real-time evaluation, but they still focus on partial capabilities, such as visual-only streaming understanding, timestamp-level response, or multi-turn dialogue, without jointly considering audio input, proactive response, and intentional silence. SOVBench fills this gap by formulating streaming omni-video understanding as a continuous audio-visual interaction task, where models not only answer when sufficient evidence appears, but also remain silent when no response is required.

4 Methodology

Task Formulation. In contrast to offline video understanding, which assumes access to a complete, fixed-length video sequence, streaming omni-video understanding operates on a continuous and potentially infinite multi-modal stream. We denote this stream as $\{x_1, x_2, \ldots, x_t, \ldots\}$, where each observation $x_t = \{v_t, a_t\}$ at time step t consists of visual frames v_t and synchronized audio segments

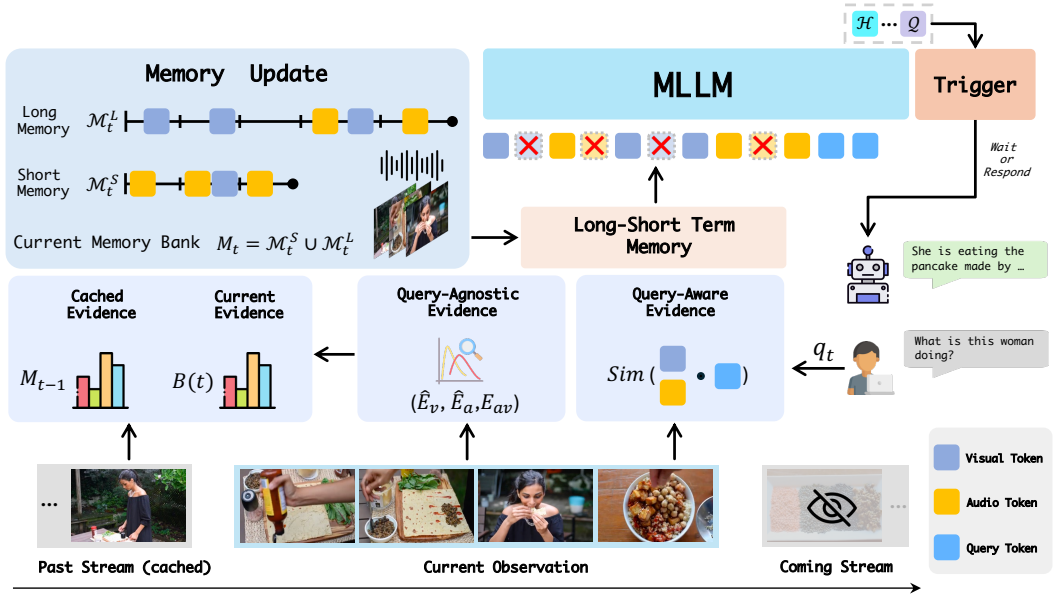


Figure 3: Framework overview of **StreamOV**. Given a omni-video stream, StreamOV constructs multimodal evidence to update a long-short term memory. The updated memory is fed into the frozen MLLM, and a lightweight hidden-state trigger decides whether to wait or generate a response.

a_t. The objective of online model ϕ is to perform online perception and interaction, generating a response y_t based on the incoming stream and a task-specific query q_t.

Due to the potential infinite horizon of streams and memory constraints, it is computationally prohibitive to retain the entire history of observations $\{x_\tau\}_{\tau=1}^{t}$ as input. To address this, we formulate the streaming process as a state transition. We define a memory state M_t to represent the historical context up to time t. This memory state is updated recursively by integrating the new observation x_t into the previous state M_{t-1}, and the output y_t is generated based on the updated memory:

$$M_t = \mathcal{F}_{mem}(M_{t-1}, x_t), \quad y_t = \phi(M_t, q_t), \tag{1}$$

where $\mathcal{F}_{mem}$ denotes the memory update function. Here, M_t serves as a compact, bounded-size summary of the accumulated stream, flexibly taking arbitrary forms such as token sequences, keyframe buffers, or other representations. Beyond information compression, a critical challenge in streaming setting is determining the optimal timing for interaction. Unlike offline systems, a streaming model must autonomously decide whether to remain silent or to respond proactively based on M_t. This necessitates identifying opportune moments when accumulated evidence is sufficient.

4.1 Multimodal Evidence Construction

To mitigate the redundancy of continuous video streaming, we propose a dual-perspective evidence construction mechanism that quantifies both query-agnostic stream dynamics and query-aware semantic relevance, allowing the online system to persist the most informative segments under a limited memory budget. We illustrate the construction process in Fig. 3.

We first extract query-agnostic characteristics to capture intrinsic stream dynamics. These include visual change S_v computed by consecutive frame differencing, audio saliency S_a derived from waveform peak detection, and audio-visual co-burst S_{cob} indicating synchronized multimodal events. Concurrently, we evaluate query-aware semantics by measuring the relevance between the query q_t and the multimodal stream. Specifically, we leverage pre-trained encoder [35, 36] to obtain the visual semantic score S_{qv} and audio semantic score S_{qa}, respectively. Since these metrics have different numerical ranges, we calculate the normalized value $S = r/r_{max}$ within current observation before aggregation, where r denotes the zero-indexed rank of the original observation.

The normalized metrics are then integrated to identify salient segments for downstream reasoning. Since audio semantic matching is often susceptible to background noise, we introduce a gating

mechanism to refine the audio evidence. The effective audio semantic score is defined as $\hat{S}_{qa} = S_{qa} \cdot \max(S_a, S_{cob})$, which prioritizes audio-query relevance only when the segment contains physically salient audio or synchronized audio-visual changes. Based on these refined metrics, we derive preliminary visual and audio evidence scores:

$$E_v = \max(S_{qv}, S_v), \quad E_a = \max(\hat{S}_{qa}, S_a). \tag{2}$$

To capture multimodal events, we further construct audio-visual aligned evidence E_{av} from two complementary cues: semantic consistency between visual and audio relevance, and event-level synchronization across modalities. Concretely, E_{av} is defined as the maximum of the semantic alignment $\min(S_{qv}, \hat{S}_{qa})$ and the event-driven burst $\min(E_v, E_a, S_{cob})$. To avoid assigning the same synchronized evidence redundantly to both modality-specific branches, we decouple the aligned evidence from the visual and audio scores:

$$\hat{E}_v = [E_v - E_{av}]_+, \quad \hat{E}_a = [E_a - E_{av}]_+, \tag{3}$$

where $[\cdot]_+ = \max(\cdot, 0)$. The final evidence is given by $\max(\hat{E}_v, \hat{E}_a, E_{av})$, and the temporal window is routed to three disentangled evidence, namely visual-only, audio-only, or audio-visual-aligned. This routing strategy provides a compact and structured evidence representation for streaming reasoning.

4.2 Long-Short Term Memory Update

Based on the constructed evidence, we update the streaming memory with a long-short term strategy. The short-term memory preserves dense observations within the current window. The long-term memory stores a sparse set of informative observations selected from both current evidence and cached historical evidence.

For each observation x_t, we determine its base evidence score $B(t)$ as the maximum value among the three routed evidence types:

$$B(t) = \max(\hat{E}_v(t), \hat{E}_a(t), E_{av}(t)). \tag{4}$$

Since $B(t)$ is derived from the rank-normalized metrics introduced in Sec. 4.1, it functions as a *local importance score* that quantifies how much an observation stands out within its own temporal neighborhood. This normalization effectively implements a local importance sampling mechanism, ensuring that salient multimodal events are identified and compared under a unified relative scale.

The streaming memory M_t is structured into a short-term memory $\mathcal{M}_t^S$ and a long-term memory $\mathcal{M}_t^L$, both of which are populated via score-based selection but differ in their temporal density. The short-term memory $\mathcal{M}_t^S$ retains the dense Top-K_S observations from the current temporal window, due to the limited duration. In contrast, the long-term memory $\mathcal{M}_t^L$ maintains a fixed Top-K_L budget of informative observations curated from the entire historical stream. As the stream progresses, $\mathcal{M}_t^L$ is dynamically updated by evaluating a candidate pool composed of the existing long-term entries and the outgoing observations from the short-term window. This results in a sparse but globally significant set of multi-modal events.

The final memory state is constructed by the union of the two buffers: $M_t = \mathcal{M}_t^S \cup \mathcal{M}_t^L$. Observations that appear in both buffers are deduplicated. All retained observations are then serialized in chronological order and fed into the base model in an interleaved visual-audio format. This memory design preserves fine-grained recent context while enabling the model to perform long-range reasoning over a curated history of salient multi-modal evidence.

4.3 Response Trigger

To achieve operational proactivity, existing approaches typically rely on explicit auto-regressive generation of specialized action tokens [5, 10, 11] or employ cascade models [12, 13]. However, the former forces a Large Language Model to explicitly generate non-semantic silence tokens, which inevitably disrupts its pre-trained distribution, while the latter delegates triggering decisions to auxiliary small models. These small routers inherently lack the deep reasoning capacity required for complex queries and fall short of the primary MLLM's comprehension, especially given the current scarcity of reliable small-scale omni-modal models.

Table 2: Evaluation on proposed SOVBench. Offline models use 1 FPS for input construction (see duration statistics in Fig. 2), whereas StreamOV sets an online budget of 64 frames.

Model	Audio	frames	SOVBench										
			Audio-Visual Context				Audio-Visual Context & QA context				SOVBench-T		
			Real-Time	Recall	Proactive	Avg.	Real-Time	Recall	Proactive	Avg.	P/R-1	P/R-0	F1
Closed-source Multimodal Models													
Gemini 2.5 Flash	✓	-	87.0	80.4	66.8	75.6	88.9	85.6	70.8	78.8	*Offline*		
Gemini 3.0 Flash	✓	-	89.5	87.6	82.6	85.6	90.3	89.7	87.2	88.6	*Models*		
Open-source Offline VideoLLMs													
Qwen2.5-VL-7B	✗	1 fps	59.7	60.8	36.0	46.8	63.8	70.1	38.9	50.6			
Qwen2.5-Omni-7B	✓	1 fps	78.4	75.3	52.1	63.9	79.5	81.4	53.6	65.5			
Qwen3-VL-30B-A3B	✗	1 fps	69.5	67.0	42.3	54.5	65.0	70.1	49.5	56.8	*Offline*		
Qwen3-Omni-30B-A3B	✓	1 fps	85.3	77.3	64.8	73.7	87.3	82.5	74.3	79.9	*Models*		
Qwen3-Omni-30B-A3B	✓	64	86.7	74.2	66.0	75.6	87.1	81.4	75.1	81.2			
OmniVinci	✓	64	59.9	62.9	38.4	48.4	65.9	74.2	44.2	54.5			
Open-source Online VideoLLMs													
Video-SALMONN S	✓	64	73.1	68.0	64.0	67.9	78.8	77.3	65.2	71.3	-	-	-
ROMA-7B	✓	1 fps	78.0	73.2	46.3	60.4	78.1	79.4	47.5	61.5	54.6/59.2	49.0/44.3	52.2
StreamOV	✓	64	86.9	73.2	78.6	**81.6**	86.7	80.4	82.1	**83.8**	**86.1/98.3**	**97.8/82.1**	**90.5**

Instead of relying on inferior external routers, we propose regarding MLLM itself as a trigger, leveraging a more intrinsic and efficient mechanism: the model's intentional readiness to respond is already encoded in the hidden states of initial decoding steps. Based on this motivation, we propose a hidden state-driven trigger that directly probes the model's internal cognitive state to make autonomous response decisions. Specifically, at any timestep t, when the model receives the accumulated memory M_t and query q_t, it initiates the prefilling and decoding process. Let $\mathcal{H}_t = \{h_{t,0}, h_{t,1}, \ldots, h_{t,k}\}$ denote the hidden states extracted from the last decoder layer, where $h_{t,0}$ is the hidden state at the final input position after prefilling, and $h_{t,i}$ for $i \geq 1$ is the hidden state at the newly generated position during the i-th decoding step. Each $h_{t,i} \in \mathbb{R}^D$ serves as a semantic latent representation of the model's current response intent. Rather than generating the full response, we utilize this leading prefix of hidden states as a compact proxy for the model's confidence and intent. To aggregate this internal cognitive evidence, we introduce a lightweight Cross-Attention Trigger module, parameterized by θ_{tr}. This module employs a learnable query vector $Q_{tr} \in \mathbb{R}^{1\times D}$ to attend to the sequence of prefix hidden states $\mathcal{H}_t$, which serve as keys and values:

$$z_t = \text{Softmax}\left(\frac{Q_{tr}\mathcal{H}_t^T}{\sqrt{D}}\right)\mathcal{H}_t, \quad p_t = \text{MLP}(z_t; \theta_{tr}), \tag{5}$$

where $z_t \in \mathbb{R}^D$ is the context-aware response intent representation, which is then fed into a classification head to yield the binary triggering logits $p_t \in \mathbb{R}^2$. The prediction p_t categorizes the current state into two actions: ***Respond*** or ***Wait***. If the accumulated multimodal evidence in M_t is insufficient to address q_t, the trigger outputs ***Wait***, and the generation process is seamlessly truncated to preserve computational resources. Conversely, if ***Respond*** is triggered, the model proceeds with full auto-regressive generation to yield y_t. During training, the trigger takes the proposed memory representation as input and we optimize the trigger θ_{tr} using a cross-entropy loss with MLLM keeping frozen. In our implementation, we only use $h_{t,0}$ for trigger prediction, which allows the triggering decision to be made immediately after prefilling and before any auto-regressive decoding step.

5 Experiments

Baseline and Benchmarks. We evaluate our method against representative Qwen-VL models, [17, 3] and Qwen-Omni [1, 2] models at both 7B and 30B scales. Since Qwen models are offline models, we further compare with state-of-the-art online streaming models under streaming scenarios. SOVBench provides two evaluation settings. The first setting directly takes the audio-visual stream as input, while the second further accumulates previous QA pairs as dialogue context for multi-round evaluation. Besides, we adopt StreamingBench [29] and OVO-Bench [28] as additional benchmarks for audio-visual QA and visual-only QA, respectively. We also report results on two offline audio-visual QA benchmarks, Video-Holmes [37] and Daily-Omni [38], to demonstrate that our method achieves superior performance beyond streaming-specific evaluation. Besides content accuracy on different benchmarks, we report category-wise accuracy on SOVBench-O. We also evaluate the silence mechanism of online models on SOVBench-T, formulating it as a binary classification task and reporting precision, recall, and F1 score to measure the model's proactive capability.

Table 3: Evaluation on online Audio-Visual and Visual-Only benchmark: StreamingBench.

Model	Audio	StreamingBench															
		Audio-Visual QA					Visual-Only QA										
		ER	SCU	SD	MA	Avg.	OP	CR	CS	ATP	EU	TR	PR	SU	ACP	CT	Avg.
Open-source Offline VideoLLMs																	
Qwen2.5-VL-7B	✗	43.6	23.2	45.2	54.0	41.5	79.0	78.1	83.9	82.7	74.5	79.8	82.4	67.1	67.4	44.0	74.5
Qwen2.5-Omni-7B	✓	48.0	22.4	50.0	70.0	49.6	80.4	79.7	81.4	80.7	76.4	81.9	83.3	68.3	66.0	43.5	74.5
Qwen3-VL-30B-A3B	✗	44.8	30.4	46.4	62.4	46.0	83.7	80.5	91.2	87.3	78.3	83.2	86.1	69.9	69.1	44.0	78.1
Qwen3-Omni-30B-A3B	✓	53.6	40.8	70.4	79.2	61.0	81.5	81.3	89.9	85.0	77.6	83.8	87.0	69.5	73.7	42.0	77.9
Open-source Online VideoLLMs																	
Flash-VStream-7B	✗	*Visual-Only Online Models Random Guessing*					25.9	43.6	24.9	23.9	27.3	13.1	18.5	25.2	23.9	48.7	23.2
VideoLLM-online-8B	✗						39.1	40.1	34.5	31.1	46.0	32.4	31.5	34.2	42.5	27.9	36.0
Dispider-7B	✗						74.9	75.5	74.1	73.1	74.4	59.9	76.1	62.9	62.2	45.8	67.6
TimeChat-Online-7B	✗						80.2	82.0	79.5	83.3	76.1	78.5	78.7	64.6	69.6	**58.0**	75.4
StreamForest-7B	✗						83.1	82.8	82.7	84.3	77.5	78.2	76.9	69.1	75.6	54.4	77.3
ROMA-7B	✓	40.4	34.8	50.4	58.8	46.1	77.0	78.1	77.9	82.1	74.8	72.9	82.4	61.8	65.9	51.1	72.4
StreamOV	✓	**63.6**	**59.2**	**79.2**	**90.4**	**68.6**	**92.6**	**87.5**	**95.0**	**90.5**	**85.7**	**94.7**	**88.0**	**80.0**	**86.1**	44.6	**86.2**

Table 4: TTFT and online E2E latency comparisons. Speedups are relative to Qwen3-Omni.

Method	Metric (ms)	20 sec	2 min	4 min	6 min	10 min
Qwen3-Omni (Baseline)	TTFT ↓	1609.1	2769.4	3526.4	4207.0	6806.8
	E2E ↓	3021.8	11091.3	17919.6	25482.9	34272.7
StreamOV (Ours)	TTFT ↓	1593.3	2385.1	2422.5	2434.5	2615.1
	E2E ↓	2453.5	4809.0	4993.2	4961.0	5350.7
StreamOV speedup vs. Qwen3-Omni	TTFT ↑	1.01×	1.16×	1.46×	1.73×	2.60×
	E2E ↑	1.23×	2.31×	3.59×	5.14×	6.41×

Table 5: Ablation study of StreamOV components.

Q-Aw.	Q-Ag.	LSMem.	Avg.
✗	✗	✗	73.7
✓	✗	✗	80.0
✗	✓	✗	78.0
✓	✓	✗	80.4
✓	✓	✓	**81.6**

Implementation Details. Our method is built upon Qwen3-Omni-30B-A3B, which contains 30B total parameters with 3B activated parameters during inference. We keep the backbone model frozen and train only a lightweight trigger module with 18.9M trainable parameters. The trigger is optimized using AdamW optimizer with a learning rate of 3×10^{-4} and a batch size of 32. The trigger training data refers to Supp. C for details. We only use **one hidden state** for trigger training. During inference, the trained trigger determines whether the model should respond or remain silent at each streaming step. We set the streaming budget to 64 frames in SOVBench and 32 frames in others. All baseline methods are sampled at 1 FPS. All experiments are conducted on 80G 8 $\times$ H100.

5.1 Evaluation on Audio-Visual Benchmarks

Both our SOVBench and StreamingBench provide multimodal question-answering evaluations. As shown in Tab. 2, SOVBench is the first multimodal multi-round QA benchmark designed for streaming omni-video understanding. The results show that SOVBench poses substantial challenges to existing omni-modal models: even Qwen3-Omni-30B achieves only 73.7% and 79.9%, whereas our StreamOV improves the performance to 81.6% and 83.8%, corresponding to a gain of 7.9% and 3.9%, respectively. Although Qwen3-Omni performs better in some scenarios, it benefits from a less constrained offline setting. Our method targets budgeted online inference and achieves superior proactive performance, a key metric in streaming settings. Furthermore, as reported in Tab. 3, StreamOV consistently outperforms audio-enabled Qwen3-Omni (+7.6%) and ROMA [26](+22.5%) on StreamingBench omni subset, achieving higher quantitative metrics in Audio-Visual QA setting.

Table 4 compares StreamOV with the Qwen3-Omni baseline in terms of time to first token (TTFT) and online end-to-end (E2E) latency across streaming video durations. With a fixed 64-frame budget, StreamOV achieves an E2E speedup of $1.23\times$–$6.41\times$ across the evaluated durations. For videos lasting 2–10 minutes, its E2E latency remains within 4.81–5.35 seconds, whereas the baseline latency increases from 11.09 to 34.27 seconds, demonstrating improved scalability to longer streams. Additional comparisons with ROMA and peak memory measurements are provided in the appendix.

Offline multi-modal benchmark results are shown in Tab. 6, StreamOV achieves competitive results on the reasoning-oriented Video-Holmes. Moreover, StreamOV consistently obtains superior perfor-

Table 6: Evaluation on offline Audio-Visual QA benchmarks: Video Holmes and Daily-Omni.

Model	Audio	Video Holmes (32 frames)								Daily-Omni (1 FPS)								
		SR	IMC	TCI	TA	MHR	PAR	CTI	Avg.	AVA	CP	CU	ES	IF	RS	30	60	Avg.
Closed-source Multimodal Models																		
Gemini 2.0 Flash	✓	56.5	54.2	43.4	44.5	43.9	55.1	50.1	49.5	62.2	73.3	63.7	63.7	76.6	75.4	67.2	68.6	67.8
Gemini 2.5 Flash	✓	43.4	46.9	43.1	51.0	37.9	43.6	39.3	43.1	73.8	66.4	72.0	68.0	78.7	81.9	69.9	77.1	73.1
GPT-4o	✗	50.0	49.6	38.8	30.0	44.0	39.2	37.0	42.0	47.9	62.6	52.3	52.6	66.2	66.3	55.6	57.5	56.5
Open-source Offline VideoLLMs																		
Qwen2.5-VL-7B	✗	34.6	21.4	15.4	12.5	17.2	12.4	15.2	19.0	37.0	46.6	33.7	37.9	52.0	44.0	39.3	42.4	40.7
Qwen2.5-Omni-7B	✓	41.8	35.1	23.4	25.0	28.9	17.5	25.9	29.0	34.5	58.8	47.7	49.7	63.0	54.9	48.7	51.1	49.8
Qwen3-VL-30B	✗	46.6	38.4	35.5	39.0	39.8	35.1	35.6	38.8	47.5	68.7	52.3	55.9	67.5	61.1	57.3	57.3	57.3
Qwen3-Omni-30B	✓	60.6	**62.3**	48.4	35.5	**50.9**	51.0	**55.2**	52.8	63.5	71.0	62.7	61.1	80.5	77.7	70.0	65.3	67.8
Open-source Online-to-Offline VideoLLMs																		
StreamOV	✓	**64.4**	56.2	**52.4**	**44.5**	48.5	**53.6**	50.4	**53.1**	**65.6**	**71.0**	**64.3**	**62.8**	**81.2**	**79.4**	**70.9**	**67.3**	**69.3**

Table 7: Evaluation on online Visual-Only QA benchmarks: OVO-bench.

Model	#Frames	Real-Time Visual Perception							Backward Tracing				Forward Active Responding				Overall Avg.
		OCR	ACR	ATR	STU	FPD	OJR	Avg.	EPM	ASI	HLD	Avg.	REC	SSR	CRR	Avg.	
Qwen2-VL-7B	64	69.1	53.2	63.8	50.6	66.3	60.9	60.7	44.4	66.9	34.4	48.6	30.1	65.7	50.8	48.9	52.7
Qwen3-Omni-30B-A3B	64	79.9	61.5	75.9	54.5	77.2	64.1	68.8	57.9	67.6	37.8	55.1	**37.7**	70.0	62.9	**56.9**	60.3
Flash-VStream-7B	1fps	25.5	32.1	29.3	33.7	29.7	28.8	29.9	36.4	33.8	5.9	25.4	5.4	67.3	60.0	44.2	33.2
StreamForest-7B	1fps	68.5	53.2	71.6	47.8	65.4	60.9	61.2	58.9	64.9	32.3	52.0	32.8	70.6	57.1	53.5	55.6
ROMA-7B	2fps	63.1	53.2	68.1	39.3	69.3	58.2	58.5	55.9	47.3	23.7	42.3	-	-	-	-	-
StreamOV	1fps	**95.3**	**78.0**	**81.0**	**66.3**	**85.2**	**76.1**	**79.5**	**66.3**	**75.0**	**46.2**	**62.4**	35.4	**74.4**	**70.0**	56.4	**64.0**

mance on Daily-Omni. We attribute these improvements to our effective evidence routing strategy, which enables the model to better identify and preserve informative audio-visual cues.

5.2 Evaluation on Visual-Only Benchmarks

We conduct experiments on *visual-only* benchmarks, where each query requires visual understanding without audio-dependent reasoning. In Tab. 3, we report the results of StreamOV on the Visual-Only QA subset of StreamingBench, where our method achieves an improvement of 8.3% over the Qwen3-Omni baseline. As shown in Tab. 7, compared with Qwen3-Omni-30B, StreamOV improves the performance by 3.7%; compared with the state-of-the-art online model StreamForest, StreamOV further achieves a gain of 8.4%. These results demonstrate that our evidence routing strategy is not dominated by audio cues, but can effectively utilize visual evidence for visual understanding.

5.3 Ablation Study

In Tab. 5, we conduct an ablation study on SOVBench-O subset to evaluate the individual contributions of the three core components in StreamOV: Query-Aware, Query-Agnostic, and LongShort-Memory. As shown in Table 4, starting from a 73.7% Avg. baseline without any of these modules, we observe that the inclusion of either Query-Aware or Query-Agnostic features leads to a performance gain, reaching 80.0% and 78.0% respectively. Combining both modules further improves the results to 80.4%, demonstrating their complementary nature in capturing multi-faceted information. Finally, the integration of LongShort-Memory also provides performance boost, elevating the overall average to 81.6%. These results validate that each component is essential and they work synergistically to achieve superior performance in streaming scenarios. More results in Supp. G.

6 Conclusion

In this paper, we introduce a novel framework StreamOV for streaming omni-video understanding task, where models must continuously process synchronized audio-visual streams, maintain bounded long-term memory, and decide when to respond. We further construct two SOVBench variants to evaluate online multi-round audio-visual QA capability and proactive response triggering with intentional silence. StreamOV combines multimodal evidence routing, long-short term memory, and an MLLM-as-a-trigger mechanism, enabling efficient streaming reasoning without explicit silence-token generation or external routers. Experiments on SOVBench-O/T and existing online/offline benchmarks demonstrate its effectiveness among streaming omni-video, offline omni-video and visual-only understanding settings.

References

[1] Jin Xu, Zhifang Guo, Jinzheng He, Hangrui Hu, Ting He, Shuai Bai, Keqin Chen, Jialin Wang, Yang Fan, Kai Dang, Bin Zhang, Xiong Wang, Yunfei Chu, and Junyang Lin. Qwen2.5-omni technical report. *arXiv preprint arXiv:2503.20215*, 2025.

[2] Jin Xu, Zhifang Guo, Hangrui Hu, Yunfei Chu, Xiong Wang, Jinzheng He, Yuxuan Wang, Xian Shi, Ting He, Xinfa Zhu, et al. Qwen3-omni technical report. *arXiv preprint arXiv:2509.17765*, 2025.

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

[4] 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.

[5] Joya Chen, Zhaoyang Lv, Shiwei Wu, Kevin Qinghong Lin, Chenan Song, Difei Gao, Jia-Wei Liu, Ziteng Gao, Dongxing Mao, and Mike Zheng Shou. Videollm-online: Online video large language model for streaming video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18407–18418, 2024.

[6] 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)*, pages 3328–3338. IEEE, 2025.

[7] Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Tao Zhong, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. Streaming video question-answering with in-context video kv-cache retrieval. *arXiv preprint arXiv:2503.00540*, 2025.

[8] Zhenyu Ning, Jieru Zhao, Qihao Jin, Wenchao Ding, and Minyi Guo. Inf-mllm: Efficient streaming inference of multimodal large language models on a single gpu. *arXiv preprint arXiv:2409.09086*, 2024.

[9] Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, et al. Streamforest: Efficient online video understanding with persistent event memory. *arXiv preprint arXiv:2509.24871*, 2025.

[10] Jiaer Xia, Peixian Chen, Mengdan Zhang, Xing Sun, and Kaiyang Zhou. Streaming video instruction tuning. *arXiv preprint arXiv:2512.21334*, 2025.

[11] 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*, pages 13448–13459, 2025.

[12] Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. Dispider: Enabling video llms with active real-time interaction via disentangled perception, decision, and reaction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24045–24055, 2025.

[13] Shehreen Azad, Vibhav Vineet, and Yogesh Singh Rawat. Streamready: Learning what to answer and when in long streaming videos. *arXiv preprint arXiv:2603.08620*, 2026.

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

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

[16] 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, April 2024.

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

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

[19] Chao Wang, Xudong Tan, Jianjian Cao, Kangcong Li, and Tao Chen. Curvestream: boosting streaming video understanding in mllms via curvature-aware hierarchical visual memory management. In *European Conference on Computer Vision*, pages 605–623. Springer, 2026.

[20] Shiwei Wu, Joya Chen, Kevin Qinghong Lin, Qimeng Wang, Yan Gao, Qianli Xu, Tong Xu, Yao Hu, Enhong Chen, and Mike Zheng Shou. Videollm-mod: Efficient video-language streaming with mixture-of-depths vision computation. *Advances in Neural Information Processing Systems*, 37:109922–109947, 2024.

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

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

[23] Jihao Liu, Zhiding Yu, Shiyi Lan, Shihao Wang, Rongyao Fang, Jan Kautz, Hongsheng Li, and Jose M Alvare. Streamchat: Chatting with streaming video. *arXiv preprint arXiv:2412.08646*, 2024.

[24] Zhenyu Ning, Guangda Liu, Qihao Jin, Wenchao Ding, Minyi Guo, and Jieru Zhao. Livevlm: Efficient online video understanding via streaming-oriented kv cache and retrieval. *arXiv preprint arXiv:2505.15269*, 2025.

[25] Joya Chen, Ziyun Zeng, Yiqi Lin, Wei Li, Zejun Ma, and Mike Zheng Shou. Livecc: Learning video llm with streaming speech transcription at scale. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 29083–29095, 2025.

[26] Xueyun Tian, Wei Li, Bingbing Xu, Heng Dong, Yuanzhuo Wang, and Huawei Shen. Roma: Real-time omni-multimodal assistant with interactive streaming understanding. *arXiv preprint arXiv:2601.10323*, 2026.

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

[28] Junbo Niu, Yifei Li, Ziyang Miao, Chunjiang Ge, Yuanhang Zhou, Qihao He, Xiaoyi Dong, Haodong Duan, Shuangrui Ding, Rui Qian, et al. Ovo-bench: How far is your video-llms from real-world online video understanding? In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18902–18913, 2025.

[29] Junming Lin, Zheng Fang, Chi Chen, Zihao Wan, Fuwen Luo, Peng Li, Yang Liu, and Maosong Sun. Streamingbench: Assessing the gap for mllms to achieve streaming video understanding. *arXiv preprint arXiv:2411.03628*, 2024.

[30] Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. In *International Conference on Learning Representations*, volume 2025, pages 42522–42556, 2025.

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

[32] Miquel Farré, Andi Marafioti, Lewis Tunstall, Leandro Von Werra, and Thomas Wolf. Finevideo. `https://huggingface.co/datasets/HuggingFaceFV/finevideo`, 2024.

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

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

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

[36] Benjamin Elizalde, Soham Deshmukh, Mahmoud Al Ismail, and Huaming Wang. Clap learning audio concepts from natural language supervision. In *ICASSP 2023-2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5. IEEE, 2023.

[37] Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*, 2025.

[38] Ziwei Zhou, Rui Wang, Zuxuan Wu, and Yu-Gang Jiang. Daily-omni: Towards audio-visual reasoning with temporal alignment across modalities. *arXiv preprint arXiv:2505.17862*, 2025.

[39] Haibo Wang, Bo Feng, Zhengfeng Lai, Mingze Xu, Shiyu Li, Weifeng Ge, Afshin Dehghan, Meng Cao, and Ping Huang. Streambridge: Turning your offline video large language model into a proactive streaming assistant. *Advances in Neural Information Processing Systems*, 38:132332–132359, 2026.

[40] Weicai Yan, Yuhong Dai, Qi Ran, Haodong Li, Wang Lin, Tao Jin, Xing Xie, Hao Liao, and Jianxun Lian. Proact-vl: A proactive videollm for real-time ai companions. *arXiv preprint arXiv:2603.03447*, 2026.

[41] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. *arXiv preprint arXiv:2406.16852*, 2024.

[42] Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. Kangaroo: A powerful video-language model supporting long-context video input. *arXiv preprint arXiv:2408.15542*, 2024.

A Benchmark Statistics

We analyze the temporal interaction structure of SOVBench-O. Among all groups, 694 contain at least one Real-Time turn, 97 contain Recall turns, and 742 contain Proactive turns. These paradigms are not mutually exclusive: 564 groups mix multiple temporal behaviors, with Real-Time + Proactive being the most common combination. Purely single-paradigm groups are also retained, including 182 purely Real-Time groups and 45 purely Recall groups.

In terms of content diversity, SOVBench-O covers 15 top-level categories and 86 fine-grained semantic categories. The source videos span a broad range of domains, including Entertainment, Education, People & Blogs, Gaming, Science & Technology, Sports, News & Politics, and Comedy. At the fine-grained level, the benchmark includes heterogeneous content such as game highlights, documentary profiles, celebrity interviews, event livestreams, short films, music videos, film trailers, parodies, social commentary, and tutorials.

B Data Filtering

As mentioned in Sec. 3.1, to ensure the high quality of the training data, we implement a metadata-driven automated assessment pipeline. As shown in Fig. 6, we design a multidimensional scoring rubric that instructs a Large Language Model (LLM) to evaluate each video across five key axes: visual dynamism, narrative coherence, information density, audio-visual alignment, and reasoning value. By analyzing the fine-grained metadata (e.g., ASR transcripts and scene descriptions) rather than the raw video, the assessor efficiently identifies samples with strong cross-modal grounding and complex temporal logic. Only videos achieving a total score of 35 or higher (out of 50) are retained, ensuring that the final dataset is rich in instructional and reasoning value.

C Trigger Training Data

Training Data. Through the data construction pipeline described in Sec. 3.1, we can derive a large number of multi-round dialogues from FineVideo. We use the prompt shown in Fig. 7 and Fig. 8 to construct positive and negative samples, where positive samples correspond to moments that require a model response and negative samples correspond to moments where the model should remain silent. Notably, the construction logic for negative samples is significantly more stringent than for positive ones, as it incorporates multi-dimensional feasibility filters to exclude ambiguous cases such as persistent visual cues or video-boundary references. This ensures that the model learns to remain silent only when the required information is genuinely absent from both audio and visual streams. In total, we generated 5,000 video samples, consisting of 2,500 positive and 2,500 negative samples. We carefully cross-check the trigger-training set against SOVBench-T at the video level and confirm that the two sets are disjoint.

Innovations of Trigger. Dispider [12] and StreamBridge [39] use separate compact decision models with dedicated <TODO> and <ACT> tokens, respectively. StreamMind [11] uses a cognition gate initialized from shallow LLM layers and trained through <response>/<silence> generation, while Proact-VL [40] uses a <|FLAG|> token's hidden state to trigger gaming commentary. These methods differ in inputs, supervision, and streaming scenarios. StreamOV instead assesses whether accumulated evidence for a persistent query is sufficient to respond. Our premise is that an autoregressive omni-MLLM implicitly encodes response readiness in its prefill state. We train only a lightweight trigger to read out this judgment, avoiding separate decision MLLMs and dedicated action-token generation [5]. The backbone can remain frozen, with Wait decisions made immediately after prefilling to avoid unnecessary autoregressive decoding.

D Ablation stduy of Triggers

Table 8 presents the ablation study of our proposed response trigger. We evaluate the binary classification performance for the *Respond* (Class-1) and *Wait* (Class-0) actions. First, compared to the baseline Qwen3omni with our proposed trigger, our StreamOV architecture significantly improves the overall F1 score from 81.4 to 90.5 under the same setting (Hidden Num = 1, using only $h_{t,0}$), demonstrating that StreamOV provides more precise memory and better internal cognitive evidence. Furthermore,

Table 8: Comparison of different trigger architectures under the same experimental setting.

Arch.	Hidden Num	Precision-1	Recall-1	Precision-0	Recall-0	F1
Qwen3omni + Trigger	1 (last prefilling)	86.8	76.7	76.7	86.8	81.4
StreamOV + Trigger	1(last prefilling)	86.1	98.3	97.8	82.1	90.5
	2(last prefilling + first decode)	88.4	95.0	93.8	85.9	90.7
	3(last prefilling + two decode)	87.3	97.5	96.7	84.0	91.0
	Linear Trigger	84.7	96.7	95.5	80.2	88.7

Table 9: TTFT and E2E latency comparison between Qwen2.5-Omni (ROMA) and StreamOV.

Method	Metric (ms)	20 sec	2 min	4 min	6 min	10 min
Qwen2.5-Omni (ROMA)	TTFT ↓	128.8	796.5	1541.0	2421.7	4000.4
	E2E ↓	1494.5	9192.7	16307.5	23729.9	30624.3
StreamOV	TTFT ↓	1593.3	2385.1	2422.5	2434.5	2615.1
	E2E ↓	2453.5	4809.0	4993.2	4961.0	5350.7

increasing the Hidden Num from 1 to 2/3 in StreamOV yields only a marginal F1 improvement (from 90.5 to 90.7/91.0). Since incorporating an additional decoding state ($h_{t,1}$) inherently introduces the computational overhead of auto-regressive generation, relying solely on the prefilling state $h_{t,0}$ (Hidden Num = 1) achieves the optimal trade-off between triggering accuracy and computational efficiency. Replacing the proposed StreamOV Trigger with linear architecture reduce the F1 score from 90.5 to 88.7.

To further assess the stability of our trigger design, we conduct five independent experimental runs of StreamOV with the trigger. Across the five runs, Precision-1, Recall-1, Precision-0, Recall-0, and F1 are 86.5$\pm$ 1.3, 98.3$\pm$0.6, 97.8$\pm$0.8, 82.6$\pm$1.8, and 90.8$\pm$1.0, respectively, where the results are report as mean$\pm$standard deviation.

E Evaluation on Audio-Visual Offline Benchmark

We provide additional latency and memory results in Tab. 9 and 10. The official ROMA implementation uses Qwen2.5-Omni-7B with an additional lightweight *speak* head. Qwen2.5-Omni (ROMA) uses less peak memory and has lower E2E latency on the shortest input, but its E2E latency exceeds that of StreamOV for videos lasting 2 minutes or longer: at 2 minutes, it requires 9.19 seconds and 21.21 GiB, compared with 4.81 seconds and 67.98 GiB for StreamOV. ROMA also achieves lower task performance, scoring 60.4 versus 81.6 on SOVBench and 72.4 versus 86.2 on StreamingBench. Because the backbones differ, this comparison serves as a system-level reference rather than a controlled same-backbone comparison.

The SOVBench is annotated in an event-driven style: queries are issued near annotated task/event times rather than at fixed clock intervals (in particular, recall queries are issued substantially later). We conduct comparative experiments following a 20-second fixed-period polling protocol. Queries are issued whenever the polling step falls within a predefined temporal window. Fixed-period Polling achieves slightly worse performance than the event-driven approach (average accuracy 79.0 vs 81.6), which may stem from cases where polling occurs while events are unfolding, leading to insufficient accumulated evidence.

F Evaluation on Long Video Benchmarks

We further evaluate StreamOV on Video-MME under the no-subtitle setting, where no external subtitles are provided and the model must rely on visual perception and audio understanding. As shown in Tab. 11, StreamOV obtains 73.5 overall accuracy and 63.4 accuracy on the long-video subset, achieving the best performance among both offline and online VideoLLMs in our comparison. Compared with Qwen3-Omni-30B, StreamOV improves the overall score by 5.0 %. The strong performance on long videos suggests that our compact memory can retain useful historical multi-modal

Table 10: Peak memory usage across video durations (GiB, lower is better).

Method	0.3 min	2 min	4 min	6 min	10 min
Qwen2.5-Omni (ROMA)	17.34	21.21	24.83	28.32	33.83
Qwen3-Omni (Baseline)	59.75	63.47	66.95	70.78	77.47
StreamOV	62.67	67.98	68.00	68.02	68.06

Table 11: Evaluation on Video-MME. We report our results without using subtitles.

Model	**#Frames**	**Video-MME**	
		Overall	**Long**
Open-Source Offline VideoLLMs			
LongVA-7B [41]	128	52.6	46.2
LLaVA-Next-Video-7B	32	46.6	–
LLaVA-OneVision-7B	32	58.2	–
Kangaroo-7B [42]	64	56.0	46.6
Qwen2.5-VL-7B	1fps	63.2	50.4
Qwen2.5-Omni-7B	1fps	64.3	–
Qwen3-Omni-30B	64	68.5	58.1
Open-Source Online VideoLLMs			
Flash-VStream-7B	1fps	61.2	50.3
VideoLLM-online-8B	2fps	52.8	44.9
Dispider-7B	1fps	57.2	–
TimeChat-Online-7B	1fps	62.5	49.2
StreamForest-7B	1fps	61.4	–
StreamOV (Ours)	64	**73.5**	**63.4**

evidence over extended temporal contexts, while preserving the efficiency required for streaming scenarios.

G Additional Ablation

Table 12: Detailed ablation study of StreamOV across different subtasks.

Q-Aw.	**Q-Ag.**	**L-Mem.**	**Real-time**	**Proactive**	**Recall**
✓	✗	✗	87.62	76.34	57.14
✗	✓	✗	86.67	72.52	64.29
✓	✓	✗	86.67	77.10	64.29
✓	✓	✓	**88.57**	**77.86**	**64.29**

To further investigate the contribution of each component, we report the accuracy on Real-time, Proactive, and Recall subtasks under 60 second accessible streaming observations. As shown in Tab. 12, the Q-Aw.-only variant performs better on Real-time questions than the Q-Ag.-only variant, while combining both evidence sources improves Proactive performance. Adding LSMem. further improves Real-time and Proactive accuracy, while Recall remains unchanged at 64.29%. LSMem. can only reorganize retained evidence within the limited observations so that posing unchanged Recall accuracy in 60-second look-back horizon (increasing accessible streaming observation to 120 seconds improves Recall accuracy to 73.2). Overall, the complete ablation study results validate the effectiveness of the proposed components.

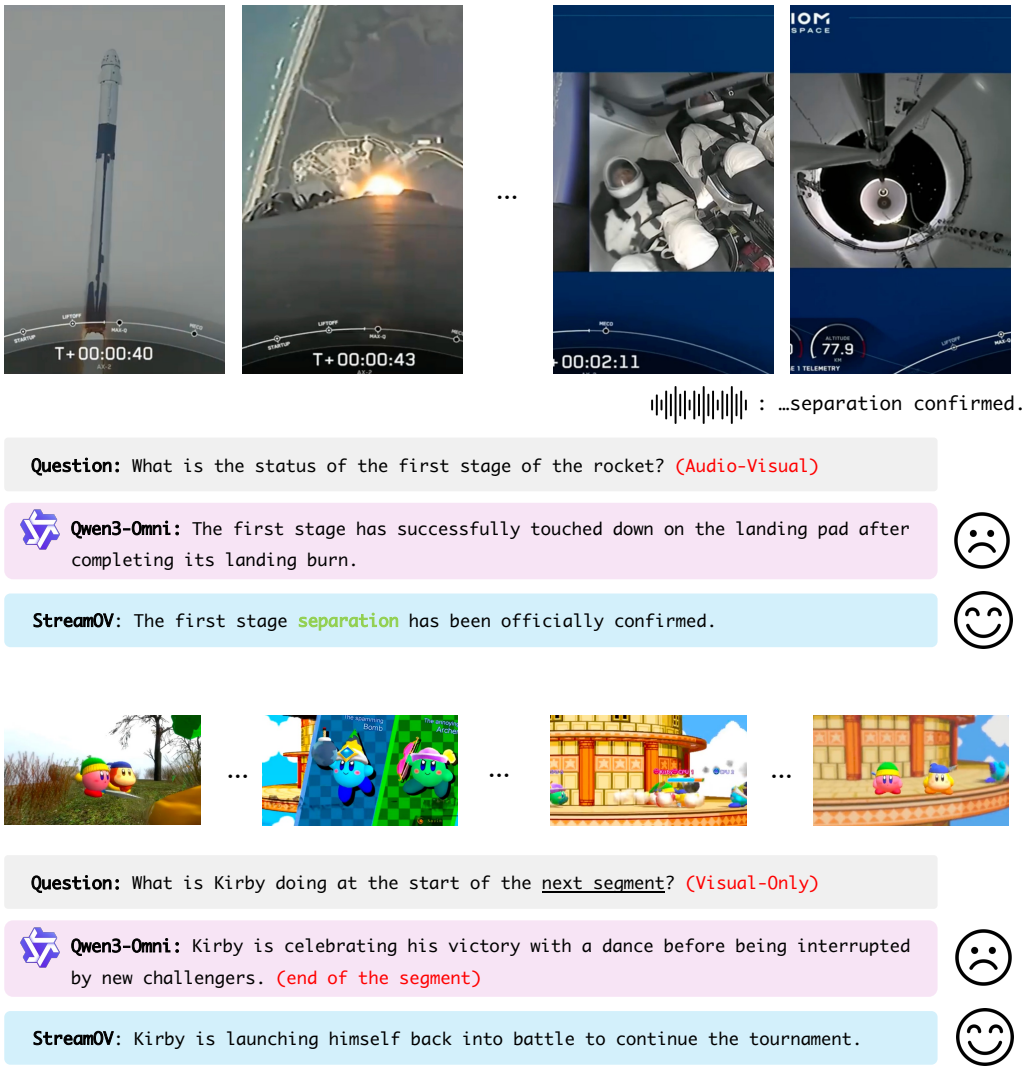


Figure 4: Qualitative cases on streaming audio-visual and visual-only understanding.

H Case Study

As shown in Fig. 4 and Fig. 5, StreamOV produces more temporally grounded and modality-aware responses than Qwen3-Omni. In the rocket launch and Kirby examples, Qwen3-Omni answers with temporally mismatched evidence, either hallucinating a later landing event or relying on the previous segment, while StreamOV correctly follows the current audio-visual context and the transition to the next segment. The lower cases further show that StreamOV can adapt to different modality requirements: it grounds the rapper's location in visual evidence while using the audio cue, and accurately captures the audio-only pronunciation instruction. These examples suggest that StreamOV better aligns its responses with the relevant streaming evidence across audio-visual, visual-only, and audio-only scenarios.

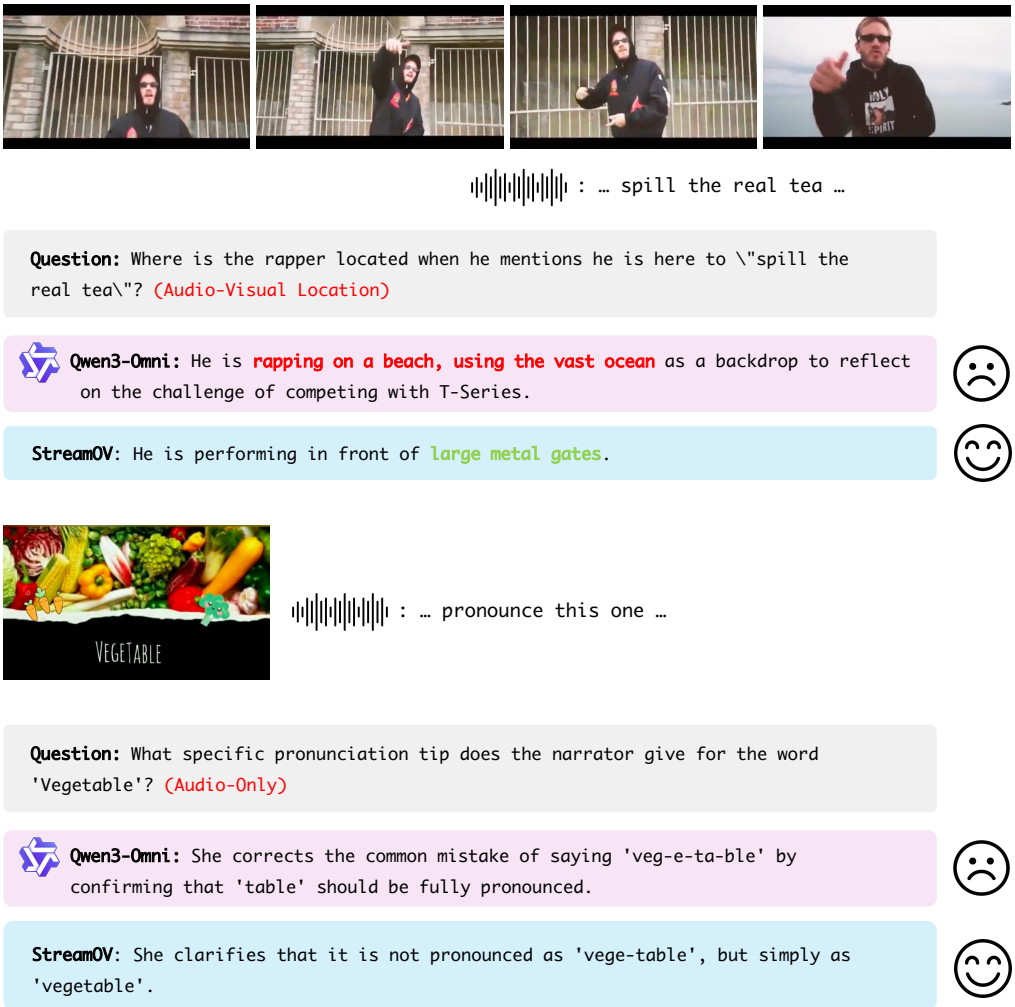


Figure 5: Qualitative cases under different modality requirements.

I Failure Scenarios of Trigger

In our experiments, for samples where visual and audio evidence is either sufficiently informative or entirely absent, the model achieves strong performance in determining *when* to respond, benefiting from the inherent reasoning capability of the MLLM and our MLLM-as-a-trigger design. Meanwhile, we also observe that when the audio-visual information is highly sparse, the trigger may become less effective due to the intrinsic limitations of the underlying MLLM.

J Limitation and Future Works

Although StreamOV provides an effective framework for streaming omni-video understanding, several aspects remain worth further exploration. First, our current memory update strategy is still based on predefined evidence scores and heuristic temporal refinement. While this design is efficient and interpretable, future work could develop more adaptive or agentic memory mechanisms that dynamically plan what to observe, store, and retrieve according to evolving user intent. Second, extending the trigger toward more fine-grained actions, such as asking clarification questions, delaying response, or actively seeking additional evidence, would make streaming interaction more flexible. Finally, SOVBench covers diverse online audio-visual scenarios, but it is still a finite benchmark built

Prompt for Data Quality Assessment

You are a Data Quality Assessor for multi-modal LLM training. Given **metadata only**, score the video for training value.
Input JSON (current schema)

- `video_uid`: string
- `content`: { `description`, `qAndA`, `scenes`: [{ `activities`, `narrative`, `props`, etc. }] }
- `ASR`: list of audio segments as { `end`: seconds, `text`: string }.

Scoring (each 0-10, integer) Use the rubric below. Choose the **best-matching band** and pick a precise score.
1) visual_dynamism (0-10): How visually rich/changing the video is.

- **0-2**: static setup; talking head; minimal motion; no meaningful props.
- **5-6**: multiple actions/scenes; moderate prop interaction.
- **9-10**: highly dynamic; complex interactions; strong visual variety.

2) narrative_coherence (0-10): Consistency and logical ordering.

- **0-2**: fragmented; contradictory; scenes don't connect; ASR unintelligible.
- **7-8**: clear temporal flow with causes/goals; strong continuity.

3) information_density (0-10): Amount of specific, non-trivial information.

- **0-2**: mostly filler; few concrete entities; repetitive.
- **9-10**: extremely dense; frequent actionable details; strong training value.

4) av_alignment (0-10): Alignment between ASR and visual narrative.

- **0-2**: unrelated (generic voiceover vs visuals); frequent mismatches.
- **9-10**: ASR tightly tracks visual actions with clear temporal grounding.

5) reasoning_value (0-10): Value for audio-visual reasoning (causal/temporal).

- **0-2**: trivial restatement; hard to form non-trivial reasoning QAs.
- **9-10**: dense event graph; abundant grounded entities; supports multi-step reasoning.

Task Return **ONLY** a valid JSON object in this schema:

```
{
"video_uid": "",
"scores": { "visual_dynamism": 0, "narrative_coherence": 0, "information_density": 0,
"av_alignment": 0, "reasoning_value": 0, "total_score": 0 },
"reasoning": { "visual_analysis": "", "content_analysis": "", "alignment_analysis": "",
"reasoning_value_analysis": "" },
"recommendation": "Keep"
}
```

Constraints: `total_score` = sum of five scores. `recommendation` is "Keep" if `total_score` $\geq$ 35, else "Discard".
Input JSON
{{PASTE_RAW_METADATA_HERE}}

Figure 6: System prompt used for Metadata-based Data Quality Assessment.

from curated videos. Scaling it to broader real-world streams with noisier environments and more open-ended user behaviors will further strengthen the evaluation of streaming omni-video systems.

K Broader Impact

This work advances streaming omni-video understanding, which has significant positive potential in real-time applications such as assistive technologies for human-robot interaction. By enabling multi-modal comprehension, it can improve situational awareness in real-world environments. However, like all large-scale multi-modal models, it may inherit biases from training datasets, leading to disparate performance across different demographics. To mitigate such risks, we should explore techniques for enhancing interpretability and controllability of streaming omni-video models in safety contexts.

L Disclosure of LLM Usage

We disclose the use of Large Language Models (LLMs) in the following aspects of this work. (1) Model Architecture: We adopt the open-source **Qwen3-Omni** as the core backbone of our proposed model for streaming omni-video understanding. (2) Data Processing: LLMs were utilized to process and refine existing open-source datasets. (3) LLMs were used to polish the language, improve grammatical accuracy. All technical contributions and final revisions were conducted and verified by the human authors.

Prompt for Trigger Sample Construction

Role: You are a professional Multimodal Data Engineer. Your task is to extract exactly ONE Positive (Trigger) training sample from a multi-turn video conversation JSON.
Rules:

1. **Candidate Selection:**
   - Pick ONE QA pair (one user message + the immediately following assistant message) from the SECOND HALF of the "conversation" list.
   - Skip any messages with category "Proactive".
   - Any QA with category "Real-time" or "Recall" is acceptable.
2. **Construction Details:** Let T = the float `time` of the selected QA pair. Let T_prev = the `time` of the most recent ASSISTANT message that occurs strictly BEFORE the selected QA's user message (or 0.0 if none). Keep all float precision to 3 decimal places.
   - Assistant content: "<Yes> " followed by the original assistant answer text.
   - Video path: "{video_uid}_0.00_{T}.mp4"
3. **Field Requirements:**
   - "summary": A concise natural-language summary of all key information from conversation turns that occur BEFORE the selected QA pair.
   - "messages": A list with exactly TWO items: user question (prefixed with "<video> ") and assistant response ("<Yes> " + original answer).
   - "videos": A list containing one string — the video path using T.
   - "t_prev", "t_selected": Audit fields (REQUIRED).
4. **Output Format (STRICT):** ONE JSON object only. No markdown fences, no explanation.

Example Output:

```
{"summary": "The video previously showed the presenter introducing a murder
mystery game...", "messages": [{"role": "user", "content": "<video> To
recap... "}, {"role": "assistant", "content": "<Yes> The detective was
Kakashi..."}], "videos": ["0zlxwPIWAqE_0.00_54.0.mp4"], "t_prev": 54.0,
"t_selected": 54.0}
```

Please process the following JSON data:

Figure 7: Prompt for extracting positive trigger samples from raw conversation data.

Prompt for Negative (Silence) Sample Construction

Role: You are a professional Multimodal Data Engineer. Your task is to extract exactly ONE Negative (Silence) training sample from a multi-turn video conversation JSON, OR explicitly skip if no clean Negative can be constructed.
Rules:

1. **Hard Selection Rules:**
   - User `time` MUST be in "valid_user_times".
   - Assistant's category MUST be "Real-time".
   - NEVER select "Recall" or "Proactive" categories.
2. **Feasibility Filter:** Reject candidates with these failure modes:
   - (F1) Continuous state (e.g., emotional state throughout).
   - (F2) Video-boundary reference (e.g., "at the start", "close out").
   - (F3) Omnipresent topic (subject permeates the entire video).
   - (F4) Persistent visual cue (answer shown by unchanging visual prop).
   - (F5) "previous-topic" window is too short or empty.
3. **Skip Mechanism (REQUIRED):** If no candidate is valid, output:
   `{"skip": true, "reason": "<short explanation>"}`
4. **Construction Details:**
   - Assistant content: exactly "<No>" (no extra text).
   - Video path: "{video_uid}_0.00_{T_prev}.mp4" (T_prev as placeholder).
   - Round all float times to 3 decimal places.
5. **Field Requirements:** Include "summary", "messages" (user with "<video> ", assistant with "<No>"), "videos", "t_prev", and "t_selected".

Example Output (Sample):

```
{"summary": "The narrator is reciting a poem...", "messages": [{"role":
"user", "content": "<video> What does the narrator say about the
soil?"}, {"role": "assistant", "content": "<No>"}], "videos":
["9xSyRhe-qOk_0.00_51.444.mp4"], "t_prev": 51.444, "t_selected": 55.055}
```

Example Output (Skip):

```
{"skip": true, "reason": "All candidates reference video boundaries; no
clean negative window."}
```

Output Format: ONE JSON object only. No markdown fences, no explanation.
Please process the following JSON data:

Figure 8: Prompt for constructing negative (silence) samples, including strict feasibility filtering and skip mechanisms.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract clearly claims that we propose a novel streaming omni-video understanding framework, StreamOV, which is consistent with main paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We have discuss limitations in Sec. J.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [N/A]

Justification: This paper does not involve theoretical result.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We have provide experimental details in Sec. 5. The dataset curation process and prompts are described in Sec. 3.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: We will release our benchmark and code to public.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Please refer to Sec. 5. We have specified the training and test details.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: We cannot provide error bar for some competitors. Besides, the current results are sufficient to show the efficacy of our method.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: Please refer to Sec. 5.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our research focuses on streaming omni-video understanding, which totally conforms the Ethics Code.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Please refer to the Sec. K in the supplementary materials.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The models and datasets used in this work do not pose a high risk for misuse.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: We cite all used public datasets and pretrained models.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: We provide comprehensive documentation for the new benchmark SOVBench in main paper and supplementary materials. We will release them to public.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: We only use objective experiments in our paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: We only use objective experiments in our paper.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [Yes]

Justification: We deseribe the usage of LLMs in Sec. L.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.