StreamReady: Learning *What* to Answer and *When* in Long Streaming Videos

Anonymous CVPR submission

Paper ID *****

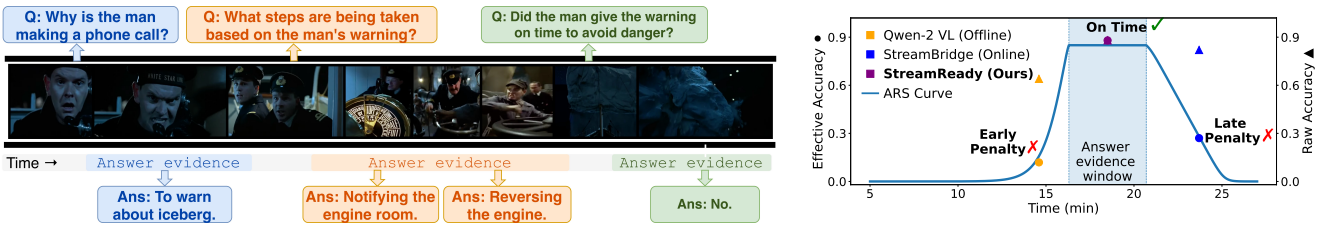

Figure 1. **Readiness-aware streaming video understanding.** *Left:* In proactive streaming settings, questions can precede their supporting evidence, requiring the model to monitor the evolving video and answer once the relevant cues appear. *Right:* Under our readiness-aware formulation, effective accuracy jointly reflects answer correctness and timing via the Answer Readiness Score (ARS). Although all models achieve similar raw accuracy on this example, ARS reveals sharp performance drops for early (hallucinatory) or late (delayed) answers. In contrast, StreamReady responds within the evidence window, preserving high effective accuracy by answering at the appropriate moment.

Abstract

*Streaming video understanding often involves time-sensitive scenarios where models need to answer exactly when the supporting visual evidence appears: answering before the evidence reflects speculation, answering after it has passed reduces real-time utility. To capture this behavior, we introduce a readiness-aware formulation of streaming video understanding with the **Answer Readiness Score (ARS)**, a timing-aware objective with asymmetric early and late penalties. When combined with correctness, ARS defines an effective accuracy that measures not just whether a model is right, but whether it answers at the appropriate moment. Building on this formulation, we introduce **StreamReady**, a framework to unify temporal reasoning with on-time answering through a lightweight readiness mechanism that decides if sufficient evidence has been observed before responding. To evaluate this capability, we further introduce **ProReady-QA**, a benchmark with annotated answer evidence windows and proactive multi-turn questions across local and global contexts. StreamReady achieves superior performance on ProReady-QA, and consistently outperforms prior methods across eight additional streaming and offline long-video benchmarks, demonstrating robust and broadly generalizable video understanding capability.*

1. Introduction

Multimodal Large Language Models (MLLMs) have significantly advanced video understanding across diverse domains [1, 2, 5, 20, 23, 38], particularly on short clips. However, their performance drops on long videos due to difficulties of reasoning over extended temporal context [59]. While recent efforts extend MLLMs to longer videos [3, 8, 14, 36, 39, 57], these models still operate offline with full-video access during inference. In contrast, streaming video understanding represents an online variant of long-video reasoning, where frames arrive sequentially, evidence may appear before or after a question, and the model must operate without ever seeing the full video. This online capability is essential for real-world settings such as surveillance, sports analytics, robotics, and assistive systems that demand timely and context-aware reasoning.

Building on advances in long-video reasoning, recent studies have begun exploring streaming video understanding [9, 18, 32, 37, 43, 47, 52]. However, most existing works focus on past-dependent (causal) reasoning where evidence is already available at question time, making answer correctness the main objective; leaving timing largely unexplored. In contrast, Many real-world scenarios require future-dependent (proactive) reasoning, where the question appears *before* the supporting evidence, requiring the model to watch the unfolding video to determine when enough information has appeared (Figure 1, *left*). In such proactive settings, models must prioritize answer timing as much as correctness: responding too early, even if correct, indicates unsupported speculation; responding too late causes unwanted delay. Developing this ability to identify the right moment to respond based on supporting evidence is therefore essential for truly effective streaming understanding.

Recently models have begun exploring such proactive

behavior by deferring responses through auxiliary MLLMs [37] or prompt-based cues [51], though at the cost of non-determinism or added compute. To evaluate this behavior, benchmarks [26, 27, 41, 42] include proactive scenarios to encourage models to wait until relevant information appears. However, they lack annotated answer evidence durations, making it difficult to verify if responses are given at an appropriate time. Consequently, these developments offer only a partial view of timing behavior since models may wait, but they lack any criteria for determining whether their chosen answer time is supported by the actual evidence.

To address these limitations, we formalize readiness-aware streaming video understanding, where the goal is not only to produce the correct answer, but to do so precisely *when* sufficient evidence appears. At its core is the **Answer Readiness Score (ARS)**, a timing-aware evaluation metric with asymmetric penalties: a harsher early penalty discourages unsupported guesses before evidence, and a milder late penalty tolerates slight delays after the evidence ends (Figure 1 *right*). Together, these penalties yield an effective accuracy that captures both answer correctness *and* timing. Building on this formulation, we propose **StreamReady**, a framework that unifies temporal reasoning with explicit answer-timing. Instead of relying on heavy auxiliary models or prompt heuristics, StreamReady introduces a lightweight learnable readiness token within its reasoning module, allowing the model to assess from its internal memory when sufficient evidence has appeared. A small readiness head monitors this token, prompting the model to answer only when appropriate, ensuring responses are both accurate and timely.

To evaluate readiness-aware understanding, we introduce **ProReady-QA**, a benchmark designed for proactive scenarios with annotated answer evidence durations and tasks covering both local and global temporal contexts, enabling systematic evaluation under ARS. Our evaluations show that StreamReady effectively bridges the gap between raw and effective accuracy, outperforming existing methods in readiness-aware streaming settings. Beyond ProReady-QA, StreamReady also achieves superior performance on other streaming benchmarks across proactive and non-proactive tasks, and generalizes well to offline long-video benchmarks. Together, our proposed formulation, method, and benchmark provide a unified foundation for advancing answer timing in streaming video understanding.

Our main contributions are as follows:

- We formalize readiness-aware streaming understanding and introduce the **Answer Readiness Score (ARS)** to jointly evaluate answer correctness and timing through asymmetric early and late penalties.
- We propose **StreamReady**, a readiness-aware framework to integrate temporal reasoning with a readiness mechanism to decide evidence sufficiency before responding.
- We develop **ProReady-QA**, a benchmark with annotated answer evidence windows and proactive multi-turn questions for evaluating timing behavior of streaming models.

2. Related Works

Offline Long Video Understanding with MLLMs. Long-video understanding aims to model extended temporal context in videos spanning from few minutes to hours. Prior works use memory-based [4, 8, 14, 36] or agent-based approaches [19, 30, 39, 57], with some adopting query-conditioned storage [3, 44]. However, these offline methods rebuild memory for every query and rely on full-video access, making them unsuitable for streaming settings where frames and questions arrive sequentially. Our work draws inspiration from query-aware conditioning but adapts it for streaming understanding without requiring memory reset.

Streaming Video Understanding with MLLMs. Inspired by long-video frameworks, memory-based [31, 32, 37, 43, 45, 50, 53] and retrieval-based approaches [9, 18, 47] extend MLLMs to streaming by processing frames online and reusing past context. While effective for answer content, they lack mechanisms for deciding answer timing; particularly important in proactive scenarios where questions precede the answer. Our readiness-aware design complements these approaches by adding explicit answer timing, avoiding premature speculation and unnecessary delays by responding precisely when the evidence appears.

Streaming Video Benchmarks. Existing streaming benchmarks [17, 24, 45, 50, 53] primarily support past-dependent question-answering, where timing has limited impact. More recent benchmarks include proactive scenarios [26, 27, 41, 42], but remain limited to short clips and local context. In contrast, ProReady-QA supports proactive reasoning over long, continuous streams with both local and global multi-turn dependencies, enabling comprehensive evaluation of on-time answering in readiness-aware streaming.

3. Method

A fundamental requirement in streaming video understanding is not only *what* a model answers but also *when* it chooses to answer. We formalize this as **readiness-aware streaming video understanding**, where a model must produce the correct answer at the appropriate moment, supported by visual evidence. Unlike existing streaming setups that evaluate only correctness and therefore cannot distinguish on-time answers from mistimed ones, our formulation explicitly accounts for both answer content *and* timing.

3.1. Framework Overview

Building on this formulation, we introduce **StreamReady**, a readiness-aware framework that learns to determine the right moment to answer by monitoring the evolving video for supporting evidence. As the video unfolds, Stream-

Ready stores them in a hierarchical memory (§3.2), retrieves and reasons over temporally relevant context when a question appears (§3.3) and uses a lightweight readiness mechanism (§3.4) to decide whether sufficient evidence is present to answer. If ready, it answers immediately; otherwise, it continues observing until the required evidence appears. Figure 2 provides an overview of the framework.

3.2. Memory Storage

Streaming video understanding can benefit from both visual and semantic history, where visual cues guide perception and prior linguistic interactions provide context, enabling efficient reuse of past information. Since streaming videos contain fine-grained details and substantial temporal redundancy, an effective system must preserve key evidence while compacting redundant content. StreamReady achieves this through two complementary memories: a Visual Memory Tree ($\mathcal{M}_{\text{V}}$) for multi-granular visual context, and a Contextual Memory Bank ($\mathcal{M}_{\text{C}}$) for long-range semantic dependencies across question–answer rounds.

Visual Memory Tree ($\mathcal{M}_{\text{V}}$). To efficiently represent long streaming videos, we maintain a multi-level Visual Memory Tree that progressively abstracts incoming frames. The lowest level $\mathcal{M}_{\text{V1}}$ stores the most recent frame embeddings in a FIFO buffer, preserving short-term details. Once full, its raw frames are compressed into a compact centroid set $\mathcal{M}_{\text{V2}} = \{c_1, c_2, \ldots, c_J\}$ via K-means clustering, forming a stable mid-level summary of the recent segment. As streaming continues, evicted frames f_o of $\mathcal{M}_{\text{V1}}$, update $\mathcal{M}_{\text{V2}}$ through EMA-based clustering with decay factor α:

$$c_j \leftarrow \begin{cases} (1-\alpha)c_j + \alpha f_o, & \text{if } \text{sim}(f_o, c_j) \geq \tau_t, \\ \text{new centroid}, & \text{otherwise} \end{cases} \tag{1}$$

where, threshold τ_t tightens in stable scenes (favoring merges) and relaxes when novelty rises (allowing new clusters), keeping $\mathcal{M}_{\text{V2}}$ compact yet adaptive. When $\mathcal{M}_{\text{V2}}$ reaches capacity J or shows distributional drift (e.g., frequent new-cluster creation), its centroids are abstracted into a coarse prototype set $\mathcal{M}_{\text{V3}} = \{s_1, s_2, \ldots s_U\}$, also updated through EMA-based clustering to support fast continuous abstraction during stable scenes.

$$s_u \leftarrow (1-\alpha)s_u + \alpha\Big(\frac{1}{|\mathcal{I}_u|}\sum_{j\in\mathcal{I}_u} c_j\Big), \tag{2}$$

If centroids become heterogeneous (e.g., low mutual similarity or persistent novelty), a lightweight mini–K-means is triggered to realign prototypes, improving coherence without full re-clustering. This hierarchical design yields a compact multi-granular memory that preserves long-range fine-grained details while reducing redundancy, enabling efficient retrieval for query-aware reasoning.

Contextual Memory Bank ($\mathcal{M}_{\text{C}}$). Beyond visual evidence, many streaming questions depend on earlier linguistic interactions. We support this with a Contextual Memory Bank $\mathcal{M}_{\text{C}}$ where each entry stores the question embedding (q_i) and the learned representation (a_i) that was used to generate its answer, forming a lightweight semantic history, that provides a complementary view to $\mathcal{M}_{\text{V}}$, enabling efficient context reuse for multi-turn reasoning.

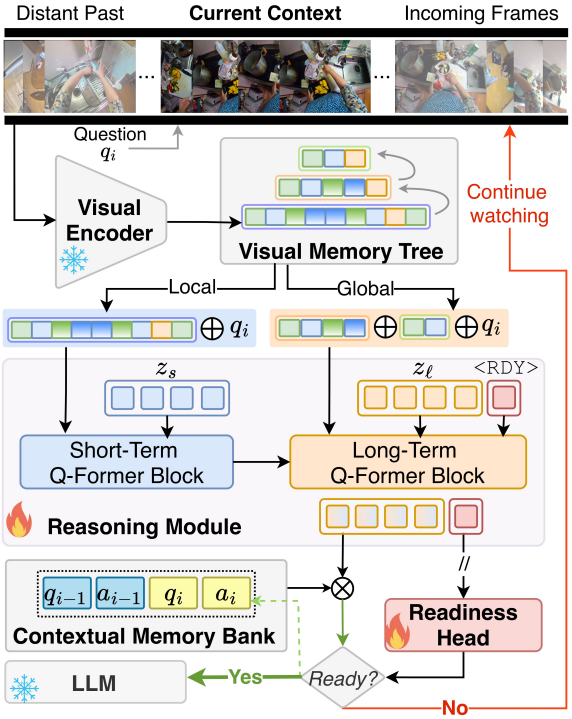

Figure 2. **Framework Overview.** StreamReady encodes streaming videos into a visual memory tree and reasons through short and long-term branches. A learnable <RDY> token, guided by a readiness head, gates the reasoning output until sufficient evidence is observed. Once ready, the long-term representation, enriched with contextual information from past QA pairs, is sent to the LLM for answering, enabling readiness-aware streaming behavior.

3.3. Query Aware Reasoning

Once a question q_i arrives, the model transitions from passive encoding to active, query-aware retrieval and reasoning. Because evidence in streaming video is distributed unevenly over time, the model must reason over both recent and older fine-grained context and abstracted summaries. To support this, we use a dual-branch Q-Former [3], consisting of a short-term branch $\mathcal{Q}_{\text{s}}(\cdot)$ that operates on $\mathcal{M}_{\text{V1}}$, and a long-term branch $\mathcal{Q}_{\ell}(\cdot)$, that operates on $\mathcal{M}_{\text{V2}}, \mathcal{M}_{\text{V3}}$.

Short-term Reasoning. To capture short-range query-relevant cues, the short-term branch $\mathcal{Q}_{\text{s}}(\cdot)$, attends to the raw frames in $\mathcal{M}_{\text{V1}}$, together with q_i, producing the short-term learned representation z_s:

$$z_s = \mathcal{Q}_{\text{s}}\big(\texttt{Concat}[\mathcal{M}_{\text{V1}}, q_i]\big), \tag{3}$$

Long-term Reasoning. In long streaming videos, relevant evidence is often buried in earlier moments, but only a small

portion might be relevant to any given query. To efficiently reason over useful long-range evidence, we perform coarse-to-fine query-aware retrieval over the higher levels of $\mathcal{M}_{\text{V}}$.

We first score all prototypes in $\mathcal{M}_{\text{V3}}$ and select the top-K high-level regions most likely to contain relevant evidence:

$$\mathcal{M}^t_{\text{V3}} = \text{Top-K}\big((W_s q_i)^\top \mathcal{M}_{\text{V3}}\big), \tag{4}$$

This acts as a semantic map lookup that highlights the most relevant video regions. The prototype scores are normalized (via softmax) to stabilize routing and sharpen focus on plausible evidence. For each selected prototype s, we then gather its associated centroids c from $\mathcal{M}_{\text{V2}}$ to form a candidate pool C_0, refine their relevance score, and select the top-m fine-grained slots:

$$\mathcal{M}^t_{\text{V2}} = \underset{c \in C_0}{\text{Top-m}} \left((W_c q_i)^\top c\right); \quad C_0 = \bigcup_{s \in \mathcal{M}^t_{\text{V3}}} \mathcal{M}^{(s)}_{\text{V2}}, \tag{5}$$

Unlike the prototype selection stage, we avoid normalizing centroid scores since sharp ranking is essential for isolating specific evidence rather than broad regions. The retrieved prototypes $\mathcal{M}^t_{\text{V3}}$, centroids $\mathcal{M}^t_{\text{V2}}$, and the short-term learned representation z_s are then passed to the long-term branch $\mathcal{Q}_\ell(\cdot)$ for reasoning:

$$z_\ell = \mathcal{Q}_\ell\big(\texttt{Concat}[\mathcal{M}^t_{\text{V3}}, \mathcal{M}^t_{\text{V2}}, q_i], z_s\big), \tag{6}$$

This two-stage retrieval mirrors episodic recall: prototypes provide coarse temporal anchors, and centroids supply the fine-grained details. Combining them with z_s enables efficient long-term reasoning that is grounded in both broad and recent evidence, producing a focused representation.

Contextual Reasoning. To effectively reuse past semantic context, we perform a contextual reasoning step over the contextual memory $\mathcal{M}_C$. The current question embedding is matched with stored previous question embeddings $\mathcal{M}_C$ to identify similar past QA interactions. A soft-gating mechanism selects the most relevant entries, and their answer representations are fused into the long-term visual feature z_ℓ through a lightweight cross-attention layer. This complements visual reasoning by incorporating prior semantic knowledge for coherent multi-turn understanding.

3.4. Readiness Mechanism

The reasoning modules determine *what* to answer, but, readiness-aware streaming understanding also requires deciding *when* to answer. To support this, we introduce a learnable <RDY> token and a Readiness Head.

Monitoring Readiness. We append the <RDY> token to the long-term reasoning representation z_ℓ within $\mathcal{Q}_\ell(\cdot)$ which already learns evidence from the visual memory tree, exposing the <RDY> token to the same evolving evidence used for answer reasoning. This ensures readiness decisions are grounded in the same representations that drive answer generation. Before the relevant evidence appears, retrieval remains weakly aligned with the question, producing a diffused, low-confidence reasoning state. As supporting evidence arrives and retrieval becomes sharper and more question-consistent; the representation shifts towards an answer-bearing state. By residing within $\mathcal{Q}_\ell(\cdot)$, the <RDY> token naturally learns to track this transition and encode how prepared the model is to answer at a given time.

A lightweight Readiness Head, monitors this token and outputs a readiness score $R_{\text{pred}} \in [0, 1]$ at each timestep. At inference, the model triggers the LLM to respond only when this score exceeds a threshold; otherwise, it continues observing future frames. When the model is ready to answer, the fused representation of z_ℓ and $\mathcal{M}_C$ is passed to the LLM for response generation, and this fused representation is stored in $\mathcal{M}_C$ as the answer representation a_i for the current question q_i. This gating enforces readiness-aware behavior by preventing premature guesses and unnecessary delays, ensuring responses occur precisely when the model judges the evidence to be sufficient.

Learning the Readiness Signal. Since the model must function online during inference but training has full-video access, we leverage this to construct weak pseudo-supervision without ever requiring ground-truth evidence timestamps. We estimate likely evidence locations by measuring similarity between the learned representation z_ℓ and centroid-level memory $\mathcal{M}_{\text{V2}}$ across time. $\mathcal{M}_{\text{V2}}$ offers a suitable balance between detail and compactness because its centroids capture fine-grained cues without the redundancy of $\mathcal{M}_{\text{V1}}$ or heavy abstraction of $\mathcal{M}_{\text{V3}}$. High-similarity centroids define a pseudo-positive temporal region P, while low-similarity ones define a pseudo-negative region N. The readiness mechanism is trained to assign higher readiness to P than to N through a pairwise contrastive loss:

$$\mathcal{L}_{ctr} = -log\ \sigma\left(R_{pred}(t^+) - R_{pred}(t^-)\right), \tag{7}$$

where $t^+ \in P, t^- \in N$. Because $\mathcal{L}_{ctr}$ alone can produce noisy, unstable readiness signals, we add a mild temporal coherence regularizer in the final objective:

$$\mathcal{L}_{rdy} = \mathcal{L}_{ctr} + \lambda_{reg} ||\nabla_t R_{\text{pred}}(t)||_1, \tag{8}$$

where $||\cdot||_1$ denotes the L1 norm. Crucially, $\mathcal{L}_{rdy}$ updates only the Readiness Head and the <RDY> token; gradients are stopped from the rest of the reasoning module so that the reasoning pathways learn *what* to answer through its standard video-text loss, while the readiness mechanism independently learns *when* to answer through the timing loss.

4. ProReady-QA Benchmark and Evaluation

To support readiness-aware streaming understanding evaluation, we introduce **ProReady-QA**, a benchmark designed specifically for long-duration streaming videos with proactive multi-turn questions and explicitly annotated answer evidence windows across local and global context (Table

Figure 3. **Examples of each task in ProReady-QA.** Here, the question and answer frames are color-coded.

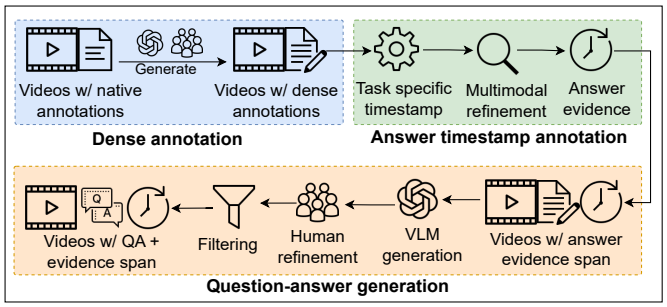

Figure 4. **Generation pipeline of ProReady-QA.**

Table 1. **Comparison of ProReady-QA with prior benchmarks.**

Benchmark	QA (K)	V Len. (min)	V Avg. (min)	Multi-turn QA Local	Multi-turn QA Global	Ans. Evidence
Non-Proactive Streaming Reasoning Benchmarks						
ODVBench [50]	6.3	<1-3	1.2	×	×	×
StreamBench [45]	1.8	2-9	4.5	×	×	×
OVBench [17]	7.0	<1-3	1.5	×	×	×
VStream-QA [53]	3.5	30-60	40	×	×	×
Proactive Streaming Reasoning Benchmarks						
ProactiveVideoQA [41]	1.4	<1-2	2.1	×	×	×
OVOBench [26]	2.8	<1-30	7.1	×	×	×
Omni-MMI [42]	2.3	<1-12	5.4	✓	×	×
StreamingBench [27]	4.5	1-10	4.1	✓	×	×
ProReady-QA	**5.0**	**30-60**	**40**	✓	✓	✓

1). ProReady-QA enables the first systematic evaluation of both answer correctness and temporal appropriateness, providing a dedicated testbed for studying answer timing behavior in streaming video models.

4.1. Task Definition

ProReady-QA spans five proactive reasoning tasks (Figure 3) requiring models to track future evidence and make temporally aware decisions: (1) **Sequential Steps Recognition (SSR):** detect process transitions; (2) **Repetitive Event Count (REC):** count recurring events; (2) **Clues Reveal Responding (CRR):** answer once certain evidence appears; (4) **Causal Trigger Detection (CTD):** detect cause–effect events; (5) **Goal-State Detection (GSD):** identify goal completion. The first three build on prior proactive tasks [26, 27], while GSD and CTD extend proactive reasoning to higher-level temporal and causal understanding.

4.2. Dataset Construction

ProReady-QA contains 10 one-hour Ego-4D [13] and 22 half-hour MovieNet [16] videos, with 5k proactive QA pairs with annotated answer evidence windows timestamps.

Sourcing Videos. We build upon long videos from VStream-QA [53], which offer diverse activities and sufficient temporal depth to support future-dependent reasoning. However, VStream-QA's questions are entirely past-dependent, making them ill-suited for proactive scenario.

Generating QA and Evidence Timestamp. We remove all past-dependent questions from VStream-QA and create new multi-turn proactive questions whose answers rely on future frames, following a semi-automatic QA generation pipeline [26, 53] with human refinement. Complex tasks (CRR and CTD) include manually authored questions to capture nuanced temporal structure. Each QA pair is annotated with precise evidence windows by aligning native annotations with multimodal cues (visual frames, subtitles, action boundaries) to identify first- and last-valid evidence timestamps. For tasks with extended events (SSR, REC), we define onset and end times to cover the full target evidence duration, yielding fine-grained temporal boundaries essential for timing evaluation. ProReady-QA also incorporates local and global multi-turn dependencies, where later questions reference earlier entities or events, reflecting real-world long-horizon reasoning. Figure 4 illustrates the overall QA generation process.

4.3. Evaluating Answer Readiness

To evaluate readiness-aware streaming performance, we introduce the **Answer Readiness Score (ARS)**, a timing-aware evaluation metric that penalizes answers given too early or too late relative to their evidence. When combined with accuracy (Acc), ARS produces an effective accuracy (Acc_e) that rewards predictions that are both correct and well-timed. For a set of N questions, ARS is defined as:

$$\text{ARS} = \frac{1}{N}\sum_{i=1}^{N}(\text{EP}_i \cdot \text{LP}_i); \quad \text{Acc}_e = \text{Acc} \times \text{ARS} \quad (9)$$

Table 2. **Performance comparison of readiness-aware streaming understanding on ProReady-QA.** †: Qwen-2-VL backbone. Metrics include accuracy (Acc.), Answer Readiness Score (ARS), effective accuracy (Acc_e). **Best** and second-best performances are highlighted.

Method	Size	SSR		CRR		REC		GSD		CTD		Average		
		Acc.	ARS	Acc.	ARS	Acc.	ARS	Acc.	ARS	Acc.	ARS	Acc.	ARS	Acc_e
Offline Video MLLMs														
InternVL2 [7]	8B	57.2	0.17	51.7	0.39	24.3	0.43	29.4	0.39	31.8	0.49	38.9	0.37	0.20
LLaVA-OneVision [21]	7B	71.8	0.48	58.2	0.53	22.3	0.45	39.4	0.24	34.8	0.18	45.3	0.38	0.29
Qwen-2-VL [38]	7B	67.9	0.32	53.1	0.39	20.7	0.31	35.1	0.52	30.3	0.28	41.4	0.34	0.20
LLaVA-NeXT-Video [55]	7B	67.4	0.47	58.3	0.41	24.3	0.42	40.3	0.31	21.8	0.38	42.4	0.40	0.31
MiniCPM-V 2.6 [49]	8B	62.7	0.56	54.2	0.62	24.5	0.35	32.7	0.43	23.4	0.35	39.5	0.46	0.23
HierarQ [3]	7B	67.4	0.32	55.8	0.45	28.7	0.44	45.2	0.38	32.8	0.43	46.0	0.40	0.27
Online Video MLLMs														
VideoLLM-online [6]	8B	53.2	0.30	51.1	0.38	18.3	0.12	30.5	0.34	24.9	0.44	29.6	0.32	0.18
Flash-VStream [53] †	7B	71.4	0.61	60.3	0.59	22.4	0.31	42.8	0.48	21.4	0.35	43.7	0.47	0.34
Dispider [32] †	7B	47.2	0.59	52.3	0.55	20.3	0.20	34.2	0.33	27.2	0.41	36.2	0.42	0.27
StreamForest [50] †	7B	65.8	0.41	57.4	0.51	24.4	0.24	49.2	0.29	32.3	0.40	45.8	0.37	0.27
StreamBridge [37] †	7B	72.2	0.72	59.7	0.65	31.9	0.43	60.3	0.57	41.4	0.49	53.1	0.60	0.42
ViSpeak [12] †	7B	67.5	0.61	55.2	0.54	25.3	0.27	48.2	0.49	27.3	0.23	44.7	0.43	0.31
InfiniPot-V [18] †	7B	69.4	0.50	60.8	0.52	34.2	0.54	58.3	0.42	37.2	0.35	52.0	0.47	0.36
StreamReady †	7B	**74.3**	**0.78**	**63.3**	**0.73**	**39.6**	**0.68**	**61.2**	**0.68**	**43.5**	**0.59**	**56.4**	**0.69**	**0.53**

where EP and LP represent early and late penalties. For each question, ProReady-QA provides a ground-truth evidence window $[t_s, t_e]$, where t_s marks when sufficient evidence first appears and t_e when it ceases to be valid. Given a model answers at time t_a, these intervals enable ARS to evaluate timing behavior through complementary early and late penalties aligned with our readiness-aware formulation.

Early Penalty (EP). Answering *before* any supporting evidence appears is the most severe readiness failure, as it reflects speculation rather than observation. To discourage such behavior, the Early Penalty sharply decreases as the model answers earlier than the evidence onset t_s, scaled by the median evidence duration τ for consistency.

$$\text{EP} = \text{softmin}\left(1, 2\,\sigma\left(\gamma_e \frac{t_a - t_s}{\tau + \epsilon}\right)\right) \tag{10}$$

Here, $\sigma(\cdot)$ is sigmoid, γ_e controls penalty sharpness, ϵ ensures numerical stability. EP $\to$ 1 as t_a approaches t_s, while early answers yield lower scores.

Late Penalty (LP). After the evidence ends at t_e, delayed responses indicate hesitation rather than hallucination. To capture this milder form of readiness error, the Late Penalty gently decreases with delay, encouraging timely responses without overly penalizing slight delays:

$$\text{LP} = \text{softmin}\left(1, \text{softmax}\left(0, 1 - \gamma_\ell \frac{t_a - t_e}{\tau + \epsilon}\right)\right) \tag{11}$$

where γ_ℓ controls decay slope, keeping LP near 1 for minor delays and lower for prolonged ones. For questions with multiple valid answers, we compute ARS separately for each turn and report their average.

5. Experiments

Datasets and Metrics. We evaluate StreamReady on ProReady-QA using accuracy, ARS and effective accuracy. To show generalization beyond readiness-aware streaming, we further evaluate on four streaming benchmarks: StreamingBench [27], OVOBench [26], OVBench [17], and VStream-QA [53]; and four offline long-video benchmarks: VideoMME [10], MLVU [58], MVBench [22], and EgoSchema [29], following each dataset's official evaluation protocol and reporting accuracy.

Implementation Details. We use Qwen-2 VL [38] as the backbone model and initialize the dual-branch Q-Former using pretrained weights [3]. For fair timing evaluation, offline models are tested on ProReady-QA by truncating videos at the evidence end and recording when the correct answer first appears, while streaming models process frames sequentially with the prompt "Answer whenever you are ready" following [26, 27]. For offline long-video benchmarks, readiness and contextual reasoning are disabled. We set $\gamma_e = 6$, $\gamma_\ell = 1$ for ARS penalties and use a readiness threshold of 0.35 for LLM trigger following [12, 37, 40].

5.1. Results

Readiness-Aware Streaming Understanding. Table 2 shows that StreamReady achieves the highest accuracy and ARS across all five ProReady-QA tasks, surpassing the best model by ~3% in accuracy and ~9% in ARS on average,

Table 3. **Performance comparison of streaming video understanding.** †: Qwen-2-VL backbone. Accuracy is the reported metric.

(a) Benchmark: StreamingBench (Strm), and OVOBench (OVO).

Method	Size	Proactive: ✓		Proactive: ×				Average	
		Strm	OVO	Strm		OVO		Strm	OVO
		Cont.	Fwd.	Real	Omni	Real	Back		
Offline Video MLLMs									
InternVL2 [7]	8B	32.4	45.4	63.7	35.8	60.7	44.0	44.0	50.0
LLaVA-OneVision [21]	7B	32.7	50.9	71.1	38.4	62.8	45.0	47.4	52.9
Qwen-2-VL [38]	7B	31.7	48.9	69.0	34.9	60.7	48.6	45.2	52.7
LongVU [34]	7B	-	48.5	-	-	57.4	39.5	-	48.5
LongVA [54]	7B	30.2	-	63.1	35.9	-	-	43.1	-
LLaVA-NeXT-Video [55]	7B	34.3	54.2	69.8	41.7	63.3	41.7	48.6	53.1
VITA 1.5 [11]	7B	27.4	53.5	52.3	33.1	63.5	41.5	37.6	52.8
MiniCPM-V 2.6 [49]	8B	35.0	-	67.4	35.0	-	-	45.8	-
HierarQ [3]	7B	35.1	48.3	69.7	44.4	67.3	48.3	49.7	54.6
Online Video MLLMs									
VideoLLM-online [6]	8B	26.6	-	36.0	28.5	20.8	17.7	30.4	19.3
Flash-VStream [53] †	7B	24.1	44.2	23.2	26.0	29.9	25.4	24.4	33.2
Dispider [32] †	7B	33.6	34.7	67.6	35.7	54.6	36.1	45.6	41.8
StreamForest [50] †	7B	-	53.5	77.3	-	61.2	52.0	-	55.6
ReKV [9] w/o offload	7B	30.7	69.1	37.4	-	-	-	-	-
StreamBridge [37] †	7B	32.6	48.4	77.0	24.1	71.3	68.1	44.6	62.6
ViSpeak [12] †	7B	43.9	54.3	70.4	61.6	66.3	57.5	58.6	59.4
InfiniPot-V [18] †	7B	-	47.9	76.4	-	65.9	47.6	-	53.8
TimeChat-Online [48]	7B	35.3	36.4	75.4	37.8	58.6	42.0	49.5	45.7
StreamAgent [46] †	7B	34.6	45.4	74.3	36.3	61.3	41.7	48.4	49.5
StreamReady †	7B	**48.2**	**58.8**	**78.3**	**63.7**	**73.6**	**72.2**	**63.4**	**68.2**

(b) Benchmark: VStream-QA. Non-Proactive.

Method	Size	RE	RM
Online Video MLLMs			
MovieChat [35]	7B	50.7	36.0
HierarQ [3]	7B	56.4	49.4
Flash-VStream [52] †	7B	57.3	53.1
ReKV [9] w/o offload	7B	55.8	50.8
InfiniPot-V [18] †	7B	57.9	51.4
StreamReady †	7B	**64.8**	**57.2**

(c) Benchmark: OVBench. Non-Proactive.

Method	Size	Avg.
Offline Video MLLMs		
LLaMA-VID [25]	7B	41.9
MiniCPM-V2.6 [49]	8B	39.1
LITA [15]	7B	20.4
InternVL2 [7]	8B	48.7
LongVA [54]	7B	43.6
LLaVA-OneVision [21]	7B	49.5
Qwen-2-VL [38]	7B	48.7
TimeChat [33]	7B	12.8
HierarQ [3]	7B	57.8
Online Video MLLMs		
VideoLLM-Online [6]	8B	9.6
MovieChat [35]	7B	30.9
VideoChat-Online [17]	4B	54.9
Flash-VStream [53] †	7B	31.2
StreamForest [50] †	7B	60.5
StreamReady †	7B	**63.9**

Table 4. **Performance comparison of offline long-video understanding.** †: Qwen-2-VL backbone. Accuracy is reported.

Method	Size	VidMME	MLVU	MVB	EgoSch
Open-source Offline Video MLLMs					
VideoChat-GPT [28]	7B	-	31.3	32.7	49.6
LLaMA-VID [25]	7B	33.2	41.9	38.5	38.5
InternVL2 [7]	8B	54.0	64.0	65.8	55.0
LongVA [54]	7B	52.6	56.3	51.3	46.7
LLaVA-OneVision [21]	7B	58.2	64.7	56.7	60.1
Qwen-2-VL [38]	7B	63.3	65.8	67.0	66.7
LongVU [34]	7B	60.6	65.4	66.9	67.6
LLaVA-Video [56]	7B	63.3	70.8	58.6	57.3
HierarQ [3]	7B	63.7	69.4	67.6	67.3
Open-source Online Video MLLMs					
MovieChat [35]	7B	38.2	25.8	55.1	53.5
Flash-VStream [53] †	7B	61.2	66.3	65.4	68.2
VideoChat-online [17]	4B	52.8	60.8	64.9	54.7
Dispider [32] †	7B	57.2	61.7	-	55.6
StreamForest [50] †	7B	61.4	70.0	70.2	-
ReKV [9] w/o offload	7B	-	68.5	-	60.7
InfiniPot-V [18] †	7B	62.8	65.8	-	65.6
VideoLLaMB [43]	7B	41.4	-	52.5	-
TimeChat-Online [48]	7B	62.5	65.4	-	-
StreamBridge [37] †	7B	64.4	69.6	64.4	66.9
ViSpeak [12] †	7B	55.0	54.1	54.1	-
StreamReady †	7B	**65.8**	**71.3**	**71.8**	**70.4**

with the largest ARS gains on REC, GSD, and CTD tasks. StreamReady's readiness mechanism reduces mistimed responses, producing tighter temporal alignment that directly lifts ARS. Additionally, its hierarchical visual memory exposes the model to clearer evidence boundaries, making readiness estimation more reliable and leading to a smaller gap between raw and effective accuracy.

Streaming video understanding. Tables 3a–3c shows StreamReady consistently outperforms prior models, achieving up to ~5% gains on proactive tasks of streaming benchmarks. While the readiness mechanism improves proactive behavior by encouraging the model to wait for sufficient evidence, most of the accuracy improvements in these benchmarks come from StreamReady's stronger evidence retrieval and long-horizon reasoning. Its visual memory offers reliable access to relevant visual cues, and the contextual memory provides complementary semantic history for multi-turn interactions. Combined with query-aware long-range reasoning, this design suppresses irrelevant historical frames and emphasizes the segments most predictive of an answer, leading to more stable and accurate predictions even when timing itself is not evaluated.

Offline long video understanding. Table 4 shows that StreamReady also maintains strong performance in offline long-video understanding, outperforming prior models despite temporal awareness being irrelevant. Here, the benefit mainly comes from its ability to integrate long-range visual structure as the memory hierarchy offers compact yet expressive summaries of extended video segments, allowing the model to reason accurately over long temporal spans.

Table 5. **Ablation studies** for each component.

Method	REC		GSD		CTD	
	Acc.	ARS	Acc.	ARS	Acc.	ARS
Contribution of each component						
Baseline	20.7	0.31	35.1	0.52	30.3	0.28
+ Trivial Reasoning [3]	28.7	0.44	48.2	0.53	34.2	0.36
+ Readiness Mechanism	28.4	0.50	47.9	0.58	34.3	0.42
+ Memory Storage	32.4	0.46	50.2	0.52	38.8	0.36
+ Query-aware Reasoning	39.4	0.48	60.9	0.53	**43.6**	0.39
+ Readiness Mechanism	**39.6**	**0.68**	**61.2**	**0.68**	43.5	**0.59**
Design choice of Readiness Mechanism						
Only MLP	39.5	0.52	60.9	0.55	42.9	0.42
LLM w/ Heuristic Triggers [51]	39.6	0.54	61.4	0.54	43.5	0.43
Auxilliary MLLM [37]	39.2	0.60	**61.2**	0.61	43.3	0.46
<RDY> + Head (Transformer)	39.5	**0.69**	**61.6**	**0.68**	43.1	0.58
<RDY> + Head (MLP)	**39.6**	0.68	61.2	**0.68**	**43.5**	**0.59**
Placement of <RDY> token						
Input of $\mathcal{Q}_s$	39.1	0.31	60.9	0.51	43.1	0.18
Learned representation of $\mathcal{Q}_s$	39.6	0.38	**61.3**	0.54	42.8	0.24
Input of $\mathcal{Q}_\ell$	39.4	0.54	61.1	0.62	**43.6**	0.49
Learned representation of $\mathcal{Q}_\ell$	**39.6**	**0.68**	61.2	**0.68**	43.5	**0.59**

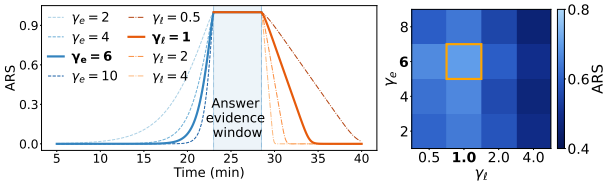

Figure 5. **Penalty sharpness for early and late responses.** *Left:* Readiness curves for different penalty strengths. *Right:* Resulting ARS, with selected γ_e, γ_ℓ combination highlighted.

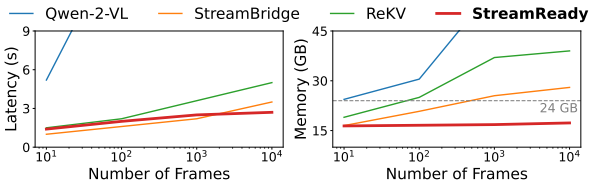

Figure 6. **Latency and memory usage analysis.**

5.2. Ablation

We ablate StreamReady's readiness mechanism on three challenging ProReady-QA tasks (REC, GSD, and CTD). Additional architectural ablations are in Supplementary.

Contribution of each component. Table 5 (*top*) shows that while adding basic reasoning [3] improves accuracy, it provides little timing benefit. Incorporating the readiness mechanism on it increases ARS modestly, but major gains occur with our stronger memory and reasoning modules. Together, these components enhance both accuracy and timing, showing that improved evidence retrieval strengthens not only *what* but also *when* the model answers.

Design choice of readiness mechanism. Table 5 (*middle*) shows that an MLP-based head is insufficient to judge readiness, while LLM or MLLM-based readiness adds only modest improvement due to weak coupling with reasoning. Embedding the <RDY> token within the reasoning module offers the best and most stable ARS improvements by directly accessing evolving evidence, with a lightweight MLP head matching Transformer performance at lower cost.

Design choice of <RDY> token placement. Table 5 (*bottom*) shows that placing <RDY> in the short-term branch yields noisy timing signals due to its limited local context, while positioning it as input to the long-term branch offers moderate but unstable results as it only passively attends to retrieved evidence. The best ARS is achieved by attaching <RDY> to the learned long-term representation, where it co-evolves with query-aligned evidence, enabling the most reliable readiness detection.

5.3. Analysis

Penalty Sharpness for Early and Late Responses. Figure 5 (*left*) shows how the sharpness parameters (γ_e, γ_ℓ) influence readiness behavior and ARS. Larger γ_e effectively suppresses premature answers, while smaller values may over-reward early guesses. Conversely, lower γ_ℓ gently tolerates slight delays, whereas higher values penalize hesitation too strongly. Figure 5 (*right*) shows a broad region of stable performance, with γ_e=6 and γ_ℓ=1 providing a balanced trade-off between avoiding speculation and allowing realistic delays, demonstrating the robustness of both StreamReady's readiness behavior and the ARS metric.

Computation Cost and Inference Latency. Figure 6 compares scalability across models as video length increases. Baseline Qwen-2-VL suffers from rapid latency growth and out-of-memory failures due to full-attention accumulation. Retrieval-based (ReKV) and activation-based (StreamBridge) methods improve efficiency but still incur rising overhead: ReKV from expanding KV caches and StreamBridge from repeated costly token-compression steps and heavy activation. In contrast, StreamReady maintains stable latency and memory by using a fixed-size compact memory of centroids and prototypes, keeping retrieval cost minimum. Its readiness mechanism, implemented with a single <RDY> token and lightweight MLP head, adds no extra inference overhead, enabling smooth scalability for long-horizon, readiness-aware streaming.

6. Conclusion

We present readiness-aware streaming video understanding, a formulation that evaluates not only *what* a model answers but also *when*, relative to visual evidence. To capture this behavior, we propose the Answer Readiness Score (ARS), a timing-aware metric with asymmetric early and late penalties. Building on this formulation, our proposed framework **StreamReady** integrates long-horizon temporal reasoning with a lightweight readiness mechanism to decide when sufficient evidence has appeared. For evaluation, we introduce **ProReady-QA**, a long streaming benchmark with proactive multi-turn questions and annotated answer evidence windows. Together, these contributions establish readiness-aware streaming as a step toward models that answer both accurately and on time.

References

[1] Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Ahmed Awadallah, Hany Awadalla, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Harkirat Behl, et al. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*, 2024. 1

[2] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: A visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 23716–23736, 2022. 1

[3] Shehreen Azad, Vibhav Vineet, and Yogesh Singh Rawat. Hierarq: Task-aware hierarchical q-former for enhanced video understanding. In *CVPR*, pages 8545–8556, 2025. 1, 2, 3, 6, 7, 8

[4] Ivana Balazevic, Yuge Shi, Pinelopi Papalampidi, Rahma Chaabouni, Skanda Koppula, and Olivier J Hénaff. Memory consolidation enables long-context video understanding. In *Forty-first International Conference on Machine Learning*, 2024. 2

[5] Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*, 2023. 1

[6] 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 *CVPR*, pages 18407–18418, 2024. 6, 7

[7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 6, 7

[8] Dingxin Cheng, Mingda Li, Jingyu Liu, Yongxin Guo, Bin Jiang, Qingbin Liu, Xi Chen, and Bo Zhao. Enhancing long video understanding via hierarchical event-based memory. *arXiv preprint arXiv:2409.06299*, 2024. 1, 2

[9] 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. In *ICLR*, 2025. 1, 2, 7

[10] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *CVPR*, 2025. 6

[11] Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Haoyu Cao, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025. 7

[12] Shenghao Fu, Qize Yang, Yuan-Ming Li, Yi-Xing Peng, Kun-Yu Lin, Xihan Wei, Jian-Fang Hu, Xiaohua Xie, and Wei-Shi Zheng. Vispeak: Visual instruction feedback in streaming videos. In *ICCV*, 2025. 6, 7

[13] Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, et al. Ego4d: Around the world in 3,000 hours of egocentric video. In *CVPR*, pages 18995–19012, 2022. 5

[14] Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *CVPR*, pages 13504–13514, 2024. 1, 2

[15] De-An Huang, Shijia Liao, Subhashree Radhakrishnan, Hongxu Yin, Pavlo Molchanov, Zhiding Yu, and Jan Kautz. Lita: Language instructed temporal-localization assistant. In *ECCV*, pages 202–218. Springer, 2024. 7

[16] Qingqiu Huang, Yu Xiong, Anyi Rao, Jiaze Wang, and Dahua Lin. Movienet: A holistic dataset for movie understanding. In *ECCV*, pages 709–727. Springer, 2020. 5

[17] 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 *CVPR*, pages 3328–3338, 2025. 2, 5, 6, 7

[18] Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. Infinipot-v: Memory-constrained kv cache compression for streaming video understanding. In *NeurIPS*, 2025. 1, 2, 6, 7

[19] Noriyuki Kugo, Xiang Li, Zixin Li, Ashish Gupta, Arpandeep Khatua, Nidhish Jain, Chaitanya Patel, Yuta Kyuragi, Yasunori Ishii, Masamoto Tanabiki, et al. Videomultiagents: A multi-agent framework for video question answering. *arXiv preprint arXiv:2504.20091*, 2025. 2

[20] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 1

[21] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024. 6, 7

[22] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *CVPR*, pages 22195–22206, 2024. 6

[23] Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, pages 237–255. Springer, 2025. 1

[24] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. 2

[25] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *ECCV*, pages 323–340. Springer, 2024. 7

[26] Yifei Li, Junbo Niu, 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? *arXiv preprint arXiv:2501.05510*, 2025. 2, 5, 6

[27] 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. 2, 5, 6

[28] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics*, 2024. 7

[29] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 6

[30] Tony Montes and Fernando Lozano. Viqagent: Zero-shot video question answering via agent with open-vocabulary grounding validation. *arXiv preprint arXiv:2505.15928*, 2025. 2

[31] Rui Qian, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Shuangrui Ding, Dahua Lin, and Jiaqi Wang. Streaming long video understanding with large language models. *Advances in Neural Information Processing Systems*, 37:119336–119360, 2024. 2

[32] 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 *CVPR*, 2025. 1, 2, 6, 7

[33] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 7

[34] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*, 2024. 7

[35] 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. 7

[36] Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. Moviechat+: Question-aware sparse memory for long video question answering. *arXiv preprint arXiv:2404.17176*, 2024. 1, 2

[37] 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. In *NeurIPS*, 2025. 1, 2, 6, 7, 8

[38] 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. 1, 6, 7

[39] Xiaohan Wang, Yuhui Zhang, Orr Zohar, and Serena Yeung-Levy. Videoagent: Long-form video understanding with large language model as agent. In *European Conference on Computer Vision*, pages 58–76. Springer, 2024. 1, 2

[40] Yueqian Wang, Xiaojun Meng, Yuxuan Wang, Jianxin Liang, Jiansheng Wei, Huishuai Zhang, and Dongyan Zhao. Videollm knows when to speak: Enhancing time-sensitive video comprehension with video-text duet interaction format. *arXiv preprint arXiv:2411.17991*, 2024. 6

[41] Yueqian Wang, Xiaojun Meng, Yifan Wang, Huishuai Zhang, and Dongyan Zhao. Proactivevideoqa: A comprehensive benchmark evaluating proactive interactions in video large language models, 2025. 2, 5

[42] Yuxuan Wang, Yueqian Wang, Bo Chen, Tong Wu, Dongyan Zhao, and Zilong Zheng. Omnimmi: A comprehensive multi-modal interaction benchmark in streaming video contexts. In *CVPR*, pages 18925–18935, 2025. 2, 5

[43] Yuxuan Wang, Cihang Xie, Yang Liu, and Zilong Zheng. Videollamb: Long-context video understanding with recurrent memory bridges. In *ICCV*, 2025. 1, 2, 7

[44] Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3272–3283, 2025. 2

[45] 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. 2, 5

[46] Haolin Yang, Feilong Tang, Linxiao Zhao, Xiang An, Ming Hu, Huifa Li, Xinlin Zhuang, Boqian Wang, Yifan Lu, Xiaofeng Zhang, et al. Streamagent: Towards anticipatory agents for streaming video understanding. *arXiv preprint arXiv:2508.01875*, 2025. 7

[47] Yanlai Yang, Zhuokai Zhao, Satya Narayan Shukla, Aashu Singh, Shlok Kumar Mishra, Lizhu Zhang, and Mengye Ren. Streammem: Query-agnostic kv cache memory for streaming video understanding. *arXiv preprint arXiv:2508.15717*, 2025. 1, 2

[48] 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. *arXiv preprint arXiv:2504.17343*, 2025. 7

[49] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. *arXiv preprint arXiv:2408.01800*, 2024. 6, 7

[50] Xiangyu Zeng, Kefan Qiu, Qingyu Zhang, Xinhao Li, Jing Wang, Jiaxin Li, Ziang Yan, Kun Tian, Meng Tian, Xinhai Zhao, Yi Wang, and Limin Wang. Streamforest: Efficient online video understanding with persistent event memory, 2025. 2, 5, 6, 7

[51] Gengyuan Zhang, Tanveer Hannan, Hermine Kleiner, Beste Aydemir, Xinyu Xie, Jian Lan, Thomas Seidl, Volker Tresp, and Jindong Gu. Avila: Asynchronous vision-language agent for streaming multimodal data interaction. *arXiv preprint arXiv:2506.18472*, 2025. 2, 8

[52] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. *arXiv preprint arXiv:2406.08085*, 2024. 1, 7

[53] Haoji Zhang, Yiqin Wang, Yansong Tang, Yong Liu, Jiashi Feng, Jifeng Dai, and Xiaojie Jin. Flash-vstream: Memory-based real-time understanding for long video streams. In *ICCV*, 2025. 2, 5, 6, 7

[54] 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. 7

[55] 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, 2024. 6, 7

[56] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 7

[57] Zhuo Zhi, Qiangqiang Wu, Wenbo Li, Yinchuan Li, Kun Shao, Kaiwen Zhou, et al. Videoagent2: Enhancing the llm-based agent system for long-form video understanding by uncertainty-aware cot. *arXiv preprint arXiv:2504.04471*, 2025. 1, 2

[58] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 6

[59] Heqing Zou, Tianze Luo, Guiyang Xie, Fengmao Lv, Guangcong Wang, Junyang Chen, Zhuochen Wang, Hansheng Zhang, Huaijian Zhang, et al. From seconds to hours: Reviewing multimodal large language models on comprehensive long video understanding. *arXiv preprint arXiv:2409.18938*, 2024. 1