VideoInsight: Rethinking Long and Streaming Video Understanding

Anonymous ACL submission

Abstract

Video understanding is moving from short-clip perception toward reliable temporal cognition. This survey reviews long and streaming video understanding through this lens: systems must locate sparse evidence, maintain memory, reason over temporal and causal structure, operate under latency constraints, and verify support. We trace evaluation from domain-specific benchmarks to long-context, multimodal, cognitive, and reliability-aware tests, highlighting benchmark-integrity risks from possible train-test overlap. We then organize offline long-video methods around the evidence bottleneck, input reduction, memory, agentic evidence search, and temporal grounding, and analyze streaming understanding as causal, bounded-memory interaction. We close with challenges for evidence-grounded, coherent, causal, robust, and trustworthy video intelligence.

1 Introduction

Video understanding is a central testbed for multimodal intelligence because videos record how objects, agents, language, sound, and environments evolve over time (Fig. 1). Unlike static images or short clips, long and streaming videos require systems to organize continuous observations into evidence, memory, and causal structure (Madan et al., 2024; Zou et al., 2024; Zhang et al., 2025a; Tang et al., 2025; Nguyen et al., 2024). Reliable systems must therefore not only recognize visible content, but also determine which moments matter, how events relate, and whether conclusions are supported by observable evidence.

This survey argues that video understanding is shifting from passive perception to reliable temporal cognition. Emerging long and streaming scenarios require models to select sparse evidence from redundant streams, preserve long-term state, reason over events and intentions, interact causally with unfolding inputs, and verify their answers. This motivates a problem-driven review that connects benchmarks, offline long-video systems, online streaming agents, and open cognitive challenges.

We organize the field around two operational paradigms. Offline long-video analysis assumes access to the complete video and emphasizes global evidence acquisition (Wang et al., 2025a); online streaming understanding processes inputs causally without future frames (Zhang et al., 2025a,b; Xia et al., 2025). Across both paradigms, evaluation has expanded from specialized recognition (Kim et al., 2025; Long et al., 2025; Xu et al., 2025c) to long-context, cross-video, multimodal, cognitive, and reliability-oriented settings (Wu et al., 2024; Wang et al., 2025c; Zhou et al., 2024), while also raising benchmark-integrity risks from reused public source videos.

For offline long-video understanding, temporal scale exposes an evidence bottleneck. Brute-force ingestion remains inefficient and can miss sparse task-relevant moments (Bai et al., 2025; Guo et al., 2025a). Existing systems therefore combine input reduction through *Token Compression* and *Keyframe Selection* (Wang et al., 2024c, 2025d; Tang et al., 2026a; Yuan et al., 2025a), *Memory-based Methods* for global coherence (Zhang et al., 2026c; Jin et al., 2025), agentic evidence search, and temporal grounding.

For online streaming understanding, the challenge is to update beliefs causally under memory and latency constraints. Systems must process incoming frames, maintain bounded state, decide when to respond, and sometimes remain silent while events unfold (Han et al., 2026; Zhang et al., 2026d). This turns Video-LLMs from retrospective analyzers into interactive agents whose correctness depends on both answer quality and response timing.

These trends point toward cognitive video intelligence: systems that reason about causes, intentions, uncertainty, counterfactual alternatives, and ab-

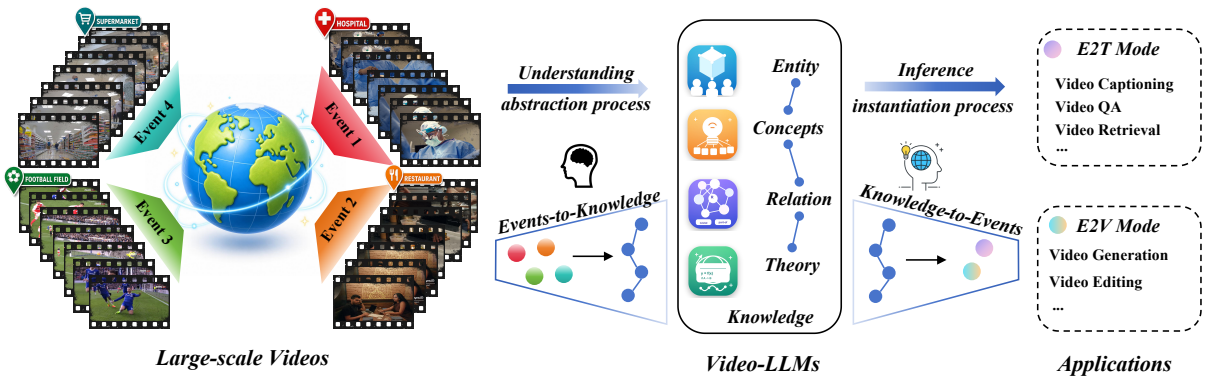

Figure 1: Motivating shift in video understanding. Video understanding connects physical observations with structured knowledge, and recent progress increasingly depends on memory, reasoning, planning, and verification rather than isolated visual recognition.

stract patterns while remaining evidence-grounded. We review the field through four connected questions: what should be evaluated, how models acquire evidence in complete videos, how they operate in causal streams, and what remains necessary for trustworthy cognitive evaluation (Fig. 2).

This study makes four main contributions:

1. We present reliable temporal cognition as a unifying perspective for long and streaming video understanding.
2. We synthesize benchmark evolution and highlight benchmark integrity as a reliability concern.
3. We analyze offline long-video methods through the evidence bottleneck: input reduction, memory, agentic search, and grounding.
4. We characterize streaming understanding as causal, bounded-memory, latency-aware interaction and connect it to cognitive video intelligence.

2 From Evaluation Landscape to Benchmark Integrity

Benchmarks define what the community treats as video understanding. Their evolution shows a shift from short-clip perception to temporal cognition: models are increasingly tested on specialized domains, social cues, long dependencies, multimodal evidence, cross-video relations, and reliability under possible data contamination.

2.1 From Short Clips to Temporal Cognition

Early benchmarks emphasized domain-specific competence in sports, medicine, advertising, virtual environments, and web interaction (Long et al., 2025; He et al., 2025b; Du et al., 2026; Yakun et al., 2025; Xue et al., 2026; Hu et al., 2025b; Jang et al., 2025; Kim et al., 2025; Zhou et al., 2025a). Such tasks remain important because real videos contain conventions, goals, and latent context that cannot be captured by action labels alone.

Human-centric, egocentric, affective, and social benchmarks push evaluation toward embodied memory, viewpoint continuity, emotions, intentions, and narrative roles (Grauman et al., 2024; Tang et al., 2026b; Jung et al., 2025; Seth et al., 2025; Yang et al., 2025d; Wang et al., 2026k; Mangalam et al., 2023; Peng et al., 2026a; Li et al., 2025j; Ye et al., 2025a; Long et al., 2025; Zhang et al., 2026e; Qiu et al., 2026b; Xie et al., 2024; Cai et al., 2025; Kong et al., 2025; Shi et al., 2025). Together, these benchmarks move evaluation from recognizing visible actions to inferring situated knowledge, embodied state, and latent mental states.

2.2 Scaling Benchmarks for Long-Video Understanding

Long single-video benchmarks such as LongVideoBench and MLVU test whether models can track events, preserve dependencies, and reason across minutes or hours of footage (Tan et al., 2025; Mangalam et al., 2023; Zou et al., 2025a; Chandrasegaran et al., 2024; Han et al., 2025; Wu et al., 2024; Wang et al., 2025c; Tao et al., 2026; Zhou et al., 2024; Ma et al., 2025a; Chen et al., 2026b; Lin et al., 2026; Zhou et al., 2025a). Their difficulty comes from sparse evidence: the decisive moment may be brief, while surrounding context remains necessary.

Evaluation has also expanded to cross-video and multimodal settings. Multi-video benchmarks require searching and synthesizing evidence across collections (Zhu et al., 2025c; Yuan et al., 2026;

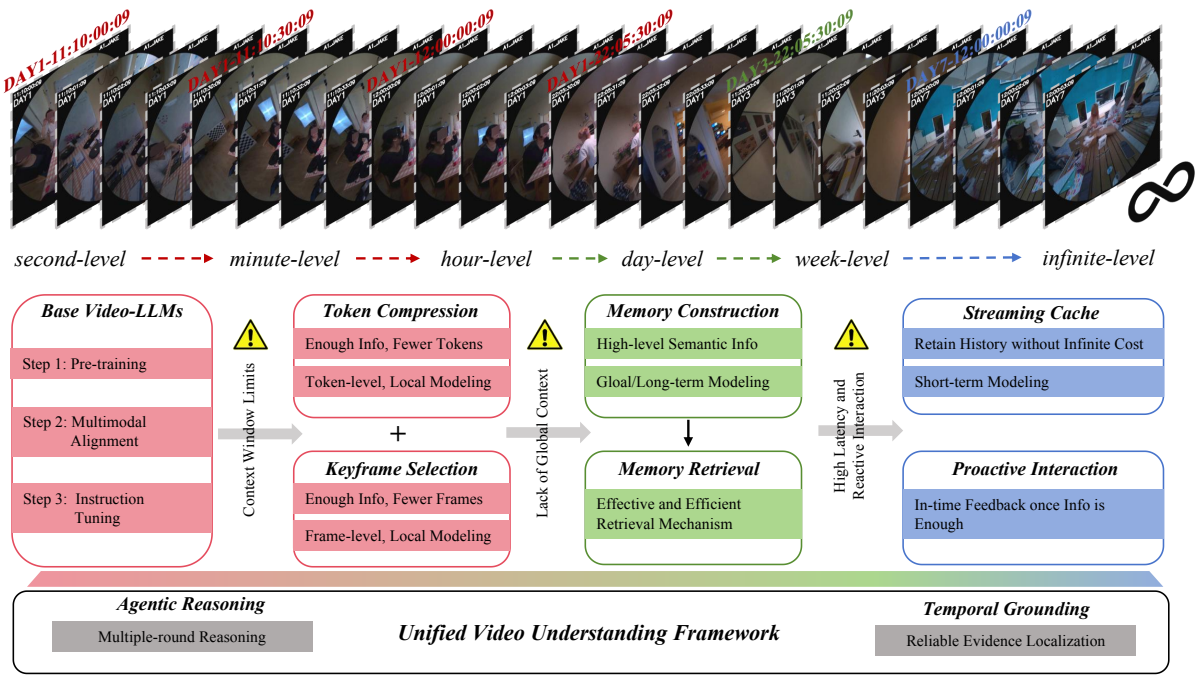

Figure 2: Overview of the proposed review framework. The survey connects evaluation, offline long-video understanding, online streaming understanding, and cognitive video intelligence through a shared trajectory from passive perception to reliable temporal cognition.

Peng et al., 2026b; Lu et al., 2025b; Hu et al., 2025d; Ye et al., 2025b; Guo et al., 2025d; Wang et al., 2026c), while all-in-one and omnimodal benchmarks combine visual, audio, OCR, subtitle, and tool-use evidence (Wang et al., 2026e; Tan et al., 2025; Zhou et al., 2024; Fang et al., 2024; Goel et al., 2026; Li et al., 2024b; Zhang et al., 2025i,g; Cheng et al., 2025b; Fu et al., 2026, 2025a; Yu et al., 2025b; Zhou et al., 2025b; Tao et al., 2026; Zhong et al., 2025; Hu et al., 2025b; Yang et al., 2025h; Hong et al., 2025). These settings motivate the evidence bottleneck: progress depends on finding, aligning, and preserving task-relevant evidence across time, videos, and modalities.

2.3 Cognitive and Reliability-Oriented Evaluation

As benchmarks become more demanding, evaluation shifts from recognizing visible content to explaining why events happen and whether answers are trustworthy. Cognitive benchmarks probe physical understanding, counterfactual inference, event evolution, and abstraction over dynamic scenes (Krojer et al., 2025; Chen et al., 2025d, 2024; Du et al., 2024; Cheng et al., 2025b; Hu et al., 2025b; Liang et al., 2025a; Yu et al., 2025b). They motivate models that ground answers in evidence, reason about causes and alternatives, and represent uncertainty.

Reliability-oriented benchmarks ask whether such reasoning is faithful, safe, and robust. They evaluate hallucination, omission, reward quality, noise resistance, and generalization under distribution shifts (Rawal et al., 2025; Wang et al., 2026h; Yu et al., 2026b; Seth et al., 2025; Lu et al., 2025a; Zhang et al., 2025d; Yang et al., 2025b; Gao et al., 2026a; Zhu et al., 2025b; Ma et al., 2025b; Zhang et al., 2025j; Meng et al., 2026; Yu et al., 2026a). Together, cognitive and reliability-oriented evaluation define temporal cognition as both inferential ability and evidence-grounded dependability.

2.4 Benchmark Integrity and Data Contamination

This expanding evaluation ecosystem also introduces a structural risk: overlap between training corpora and evaluation benchmarks. Because many Video-LLMs reuse public datasets or their source videos for instruction tuning, reported gains may partially reflect memorization, source leakage, or near-duplicate exposure rather than genuine generalization. This risk is especially serious in video understanding, where the same source video can be repackaged into QA, captioning, grounding, retrieval, and reasoning datasets.

To support reliable assessment, we manually

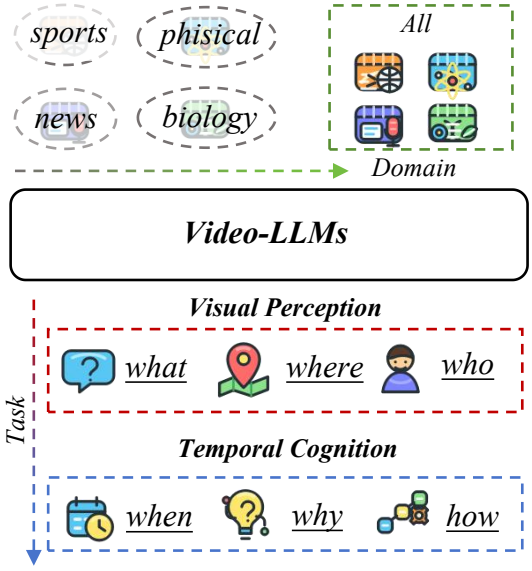

Figure 3: Evolution of video understanding benchmarks across domains and tasks. Evaluation has expanded from specialized domain recognition to human-centric, multimodal, cognitive, and reliability-oriented tasks that require temporal evidence and latent-state reasoning.

Evaluation Focus	Representative Benchmarks
Domain and task-specific	AdsQA (Long et al., 2025), AVUTBench, ExpVid (Xu et al., 2025c)
Long-context and cross-video	ALLVB (Tan et al., 2025), CVBench (Zhu et al., 2025c), EgoLifeQA (Yang et al., 2025d)
Cognitive reasoning	CounterVQA (Chen et al., 2025d), EnvQA, EventHallusion
Reliability and affective cues	AutoEval-Video, CUEBench (Yu et al., 2026b), EmoVid (Qiu et al., 2026b)

Table 1: Representative clean evaluation benchmark candidates for reliable assessment. The complete clean list appears in Appendix Table 6, while benchmarks with identified overlap are listed in Appendix Table 7.

reviewed the dataset provenance reported across hundreds of benchmark and training papers. For each evaluation benchmark, we inspected the data-source descriptions in the corresponding literature and compared them with the training-data lists reported by Video-LLM papers, marking a potential overlap whenever the same dataset, source video collection, or clearly derived benchmark appeared on both sides. This literature-based procedure is intentionally conservative: it does not claim sample-level leakage in every case, but it exposes where evaluation may be compromised by shared data origins. Based on this review, Table 1 reports a condensed clean subset, while Appendix Table 6 gives the full clean list and Appendix Table 7 lists potentially contaminated benchmarks. We recommend that researchers prioritize clean benchmarks for assessment or restructure existing datasets to develop novel evaluation standards. This framing also clarifies the role of evaluation in the rest of the survey: benchmarks reveal the capability gaps

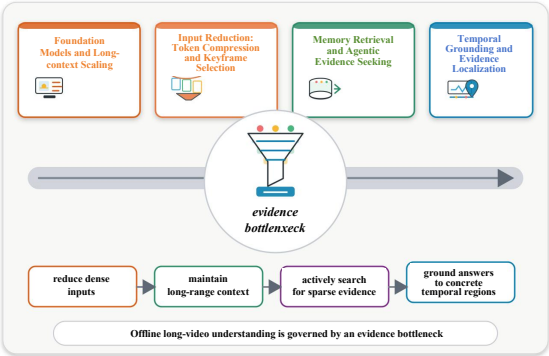

Figure 4: Offline long-video Understanding.

that offline long-video methods, online streaming systems, and cognitive video intelligence aim to close.

Takeaway. Evaluation is the entry point of the survey: the benchmark landscape shows that video understanding is no longer only about recognizing clips, but about testing temporal evidence, multimodal reasoning, cognitive abstraction, and trustworthy generalization.

3 Offline Long-Video Understanding

Offline long-video understanding analyzes complete videos that span minutes to hours. Unlike short-clip recognition, the central problem is evidence management: systems must compress dense inputs, maintain long-range state, retrieve sparse moments, and ground answers in time. We organize this section as a pipeline from foundation models to input reduction, memory, agentic search, and grounding.

3.1 Long-Context Video Foundations

Base Video-LLMs have evolved from simple encoder–LLM bridges into interleaved multimodal architectures with temporal adapters, dual-stream processing, hierarchical grounding, and longer context windows (Liu et al., 2024a; Ye et al., 2023; Li et al., 2025a; Maaz et al., 2024; Zhang et al., 2023; Li et al., 2025g; Lin et al., 2024; Jin et al., 2024; Li et al., 2024a, 2025c; Ye et al., 2024; Liu et al., 2024b; Rasekh et al., 2026; Wu et al., 2026b; Azad et al., 2025; Liang et al., 2025b; Xu et al., 2024; Liu et al., 2024c; Fei et al., 2024; Munasinghe et al., 2025). For long-video scalability, token compression, state-space backbones, and sequence parallelism extend usable context to million-token levels (Xu et al., 2025b; Wang et al., 2025f; Zhang et al., 2024b; Chen et al., 2025e; Zhao et al., 2024; Bai et al., 2025; Ren et al., 2025; Liu et al.,

2025a). Performance increasingly depends on training and reasoning strategies: multi-stage pretraining, instruction tuning, reinforcement learning, and test-time scaling shift models from language-prior reliance toward evidence-based reasoning (Zohar et al., 2025; Wang et al., 2025e; Zhu et al., 2025a; Li et al., 2024c; Guo et al., 2025a; Ren et al., 2024; Zeng et al., 2025; Vinod et al., 2025; Li et al., 2025d; Chen et al., 2026e; Yan et al., 2026c; Jiang et al., 2025b; He et al., 2025d). Yet brute-force ingestion remains costly, motivating explicit evidence reduction and memory.

3.2 Input Reduction for Sparse Evidence

Token compression (Table 2) preserves spatio-temporal evidence under tight budgets. Methods summarize streams into compact grids, prune or merge redundant tokens, and route capacity by query relevance, motion, or hierarchical structure (Zhang et al., 2024a; Li et al., 2024d; Zhang et al., 2025f; Doorenbos et al., 2025; Qi et al., 2026; Wang et al., 2025d; Zhang et al., 2025h,c; Cho et al., 2025; Sun et al., 2025a; Man et al., 2025; Sun et al., 2024a; Shen et al., 2024; Guo et al., 2025b; Faure et al., 2025; Kim and Seo, 2025; Ren et al., 2023; Liu et al., 2025c). Cache-aware and state-space designs further reuse KV memory and rebuild backbones for long sequences (Patel and Patel, 2025; Sai et al., 2026; Tao et al., 2025; Wang et al., 2024c; Huang et al., 2026b; Di et al., 2025; Li et al., 2026a; Jiang et al., 2025a; Xu et al., 2025a; Song et al., 2025a).

Keyframe selection complements compression by deciding which frames or clips enter the model. Early methods rely on CLIP matching, scene segmentation, and clip-level windows to balance relevance and redundancy (Wu et al., 2026c; Huang et al., 2026a; Liu et al., 2025b; Huang et al., 2025a; Liang et al., 2024; Zhang et al., 2025e; Hu et al., 2025a; Chen et al., 2026c; Fang et al., 2025a; Chen et al., 2026d; Sun et al., 2025b; Fu et al., 2025b; Dai et al., 2025). Recent work treats sampling as active reasoning, using tree search, self-reflection, and reinforcement learning to update beliefs across rounds (Zou et al., 2025b; Arnab et al., 2026; Hu et al., 2025c; Wang et al., 2025h; Chen, 2025; Yao et al., 2025; Tan et al., 2026a; Zhou et al., 2026b; Tang et al., 2026a). Multimodal cues (captions, subtitles, timestamps) and multi-resolution allocation further sharpen evidence localization while controlling token cost (Yuan et al., 2025a; Lin et al., 2025a; Sheng et al., 2025; He et al., 2025a; Wang et al., 2026f; Bao et al., 2025; Wang et al., 2025b). Even so, both compression and selection mainly address local evidence density and degrade on hour-scale videos (*e.g.*, LVBench in Table 2), where global memory is essential.

3.3 Memory and Agentic Evidence Search

Memory-based methods (Table 2) retain and organize information across long horizons. Sparse visual buffers, compact summaries, and language repositories store essentials with controlled cost (He et al., 2024; Song et al., 2025b, 2024; Chen et al., 2026e; Jin et al., 2025; Sasse et al., 2025; Kahatapitiya et al., 2025; You et al., 2024). Episodic and event-level memories build hierarchical narratives, while object- and environment-centric memories track spatial relations for embodied tasks (Cheng et al., 2025a; Yeo et al., 2025; Lin et al., 2025b; Santos et al., 2025; Zuo et al., 2026; Goletto et al., 2024; Huang et al., 2025b; Fan et al., 2025b; Chu et al., 2025). Memory-feedback and online updates further connect storage with compression, frame selection, and event-boundary detection (Yamao et al., 2026; Diko et al., 2025; Lan et al., 2024; Guo et al., 2026). Memory thus determines which past observations remain available for later reasoning, how evidence is indexed, and whether long-range dependencies can be recovered when needed.

Agentic methods turn retrieval into an observe–think–act loop. Active perception lets models iteratively inspect or zoom into segments, while tool-augmented and planning-based agents decompose tasks and balance global exploration with local exploitation (Wang et al., 2025i; Fan et al., 2025a; He et al., 2025c; Yin et al., 2026; Wang et al., 2026j; Ding et al., 2025; Gao et al., 2025; Tian et al., 2025; Yuan et al., 2025b; Gao et al., 2026b; Yang et al., 2025e; Qiu et al., 2026a; Yang et al., 2025g; Xie et al., 2025). Reflection, hallucination-aware feedback, and multi-agent collaboration improve verification by judging evidence sufficiency through specialized planner, observer, and verifier roles (Pu et al., 2025; Li et al., 2025f; Wang et al., 2026i; Zhi et al., 2025; Chen et al., 2025c; Zhu et al., 2026; Chen et al., 2025b; Li et al., 2025e; Zhou et al., 2026a; Yang et al., 2026c; Yan et al., 2026a; Chen et al., 2025a; Rege et al., 2026; Liu et al., 2025d). Agentic search makes evidence acquisition explicit, but its quality depends on the underlying VLM, tool reliability, and search budget; longer loops add latency and may compound errors when intermedi-

ate observations are noisy.

3.4 Temporal Grounding and Evidence Localization

Temporal grounding links outputs to timestamps, moments, or spatio-temporal regions, anchoring retrieval, reasoning, and verification. Existing work introduces explicit temporal tokens, time-aware representations, coarse-to-fine localization, and spatio-temporal entity grounding to align language with video evidence (Wang et al., 2024b, 2026g; Guo et al., 2024; Li et al., 2026b; Guo et al., 2025c; Huang et al., 2024; Li et al., 2025b; Hannan et al., 2025; Chen et al., 2026a; Shen et al., 2025; Li et al., 2025c; Gu et al., 2026; Fang et al., 2025b; Di and Xie, 2024; Xu et al.; Wang et al., 2024a). More recent methods optimize grounding as verifiable reasoning through reinforcement learning, feedback rewards, causal event modeling, and streaming cues (Luo et al., 2025; Li et al., 2025i; Xu et al., 2025d; Chen et al., 2025f; Shah et al., 2026; Cheng et al., 2026; Yu et al., 2025a; Wang et al., 2026d). Grounding therefore provides a citation mechanism for video answers: it allows claims to be traced back to evidence and supports hallucination diagnosis required by agentic verification.

Takeaways. Offline long-video understanding is governed by an evidence bottleneck:

1. Input reduction defines the efficiency floor: compression and keyframe selection cluster near baselines, helping with budgets but bounded in absolute gain.
2. Memory quality determines the long-context ceiling: memory methods show larger variance on hour-scale benchmarks.
3. Agentic integration turns modules into evidence-seeking systems by combining compression, memory, retrieval, verification, and grounding, but its effectiveness still depends on the underlying agents.

3.5 Memory, Agentic Search, and Grounding

Memory-based methods retain information across long horizons using sparse visual buffers, language repositories, event-level episodes, and object-centric state (He et al., 2024; Song et al., 2025b, 2024; Chen et al., 2026e; Jin et al., 2025; Sasse et al., 2025; Kahatapitiya et al., 2025; You et al., 2024; Cheng et al., 2025a; Lin et al., 2025b; Zuo et al., 2026; Chu et al., 2025). Agentic methods then turn retrieval into an observe–think–act loop,

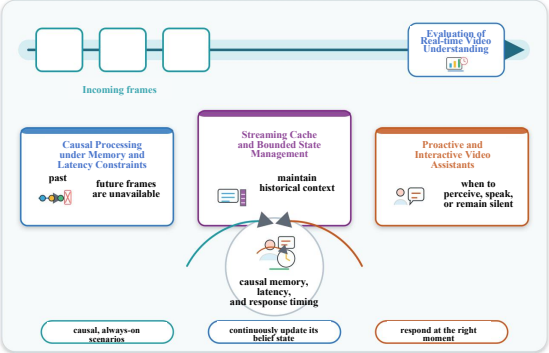

Figure 5: Online streaming video understanding.

deciding what to inspect, which tools to invoke, and how to verify evidence (Wang et al., 2025i; Fan et al., 2025a; He et al., 2025c; Wang et al., 2026j; Ding et al., 2025; Yuan et al., 2025b; Zhou et al., 2026a; Yang et al., 2026c; Yan et al., 2026a; Liu et al., 2025d). Temporal grounding anchors these processes by linking answers to timestamps, moments, entities, or regions, enabling evidence citation and hallucination diagnosis (Wang et al., 2024b, 2026g; Guo et al., 2024; Li et al., 2026b; Guo et al., 2025c; Huang et al., 2024; Li et al., 2025b; Chen et al., 2026a; Shen et al., 2025; Li et al., 2025i; Cheng et al., 2026).

Takeaway. Offline long-video systems are limited by an evidence bottleneck: input reduction improves efficiency, memory determines long-range recall, and agentic grounding makes evidence acquisition verifiable.

4 Online Streaming Video Understanding

Online streaming video understanding extends offline analysis to causal, always-on scenarios. Without future frames, models must process incoming evidence under memory and latency budgets and decide whether and when to interact.

4.1 Causal Streaming under Real-Time Constraints

Streaming systems update from bounded history rather than reprocessing the full video after the fact, which creates a trade-off among temporal coverage, memory, and latency: sparse sampling may miss brief events, dense sampling raises cache pressure, and premature or delayed responses both fail in interactive use. Evaluation must therefore consider causal evidence availability, latency, and response timing alongside answer quality.

Methods	Baseline	F/S	Video-MME (11–3600s)	LongVideoBench (15–3600s)	LVBench (∼4101s)	MLVU (180–7200s)
Proprietary and Base Models						
GPT-4o (Hurst et al., 2024)	-	-	71.9	66.7	34.7	64.6
Gemini-2.5-Pro (Comanici et al., 2025)	-	-	82.4	69.8	67.4	79.6
Seed1.5-VL (Guo et al., 2025a)	-	-	77.9	74.4	64.6	82.1
Qwen3-VL-8B (Bai et al., 2025)	-	-	79.2	-	67.7	84.3
Input Reduction						
ReTaKe (Wang et al., 2024c)	LLaVA-Video-7B	-	64.0	59.7	48.5	67.8
AdaRETAKE (Wang et al., 2025d)	Qwen2.5-VL-7B	-	67.7	62.6	51.2	75.0
AdaptToken (Qi et al., 2026)	Qwen2.5-VL	-	70.5	65.2	54.8	76.8
TSPO (Tang et al., 2026a)	Qwen2.5-VL-7B	64	65.5	64.2	46.4	74.3
DATE (Yuan et al., 2025a)	Qwen2.5-VL-7B	256	67.3	63.3	47.4	-
SeViCES (Sheng et al., 2025)	Qwen2.5-VL-7B	64	65.5	63.9	45.4	72.2
Memory and Agentic Reasoning						
VideoDetective (Yang et al., 2026a)	Qwen2.5-VL-7B	-	62.1	-	51.1	45.8
VideoMem (Jin et al., 2025)	Qwen3-VL-8B	-	73.6	63.3	58.5	77.4
VideoMind (Liu et al., 2025d)	Qwen2-VL-7B	-	58.2	-	40.8	64.4
VideoChat-A1 (Wang et al., 2026j)	Qwen2.5-VL-7B	42.0	71.8	64.2	-	71.9
Symphony (Yan et al., 2026a)	DeepSeek-R1/DeepSeek-V3/Seed1.6-VL	-	-	77.1	71.8	81.0
MR.Video (Pang and Wang, 2025)	Gemini-2.0-Flash/GPT-4o	-	61.8	61.6	60.8	-
Temporal Grounding						
TPO (Li et al., 2025i)	LLaVA-Video-7B	-	65.6	60.1	-	71.1

Table 2: Condensed reported results on long-video understanding benchmarks. F/S denotes frames per sample. Rows prioritize cited representative methods and summarize trends across proprietary/base models, input reduction, memory, agentic reasoning, and temporal grounding; the complete table is provided in Appendix Table 4. Results remain protocol-dependent and should not be read as a strict fair ranking.

Methods	Baseline	F/S	OVO	Streaming
GPT-4o (Hurst et al., 2024)	-	64	63.6	73.28
Qwen3-VL-8B (Bai et al., 2025)	-	4	81.4	80.59
WAT (Han et al., 2026)	Qwen2.5VL-7B	1fps	64.7	77.70
WeaveTime (Zhang et al., 2026d)	Qwen2-VL-7B	1fps	66.28	75.39
QueryStream (Zhang et al., 2026b)	Qwen2.5-VL-7B	1fps	63.1	75.32
SAVEMem (Wu et al., 2026a)	Qwen2.5-VL-7B	1fps	74.93	76.0
CurveStream (Wang et al., 2026a)	Qwen3-VL-8B	10–20	80.76	85.56
DSCache (Pang et al., 2026)	Qwen2.5-VL-7B	1fps	71.7	82.32
AURA (Lu et al., 2026)	Qwen3-VL-8B	-	79.8	83.2

Table 3: Condensed streaming benchmark results. F/S denotes frames per sample or frame rate; the full table is in Appendix Table 5.

4.2 Bounded Memory and Cache Management

Streaming caches keep historical context under fixed memory budgets. Foundational designs treat KV caches as persistent memory, storing hierarchical caches as reusable substrates for long-term recall (Zhang et al., 2026a; Ning et al., 2025; Di et al., 2025), while compression methods such as InfiniPot-V and VideoScan prune redundant tokens or use semantic carriers to stabilize memory during continuous encoding (Kim et al., 2026b; Yang et al., 2025f; Li et al., 2025h). More advanced systems integrate retrieval with compression, route frames by semantic trajectories or action boundaries, and decouple historical caches from immediate context to balance long-term retention with recent events (Patel and Patel, 2025; Agarwal et al., 2026; Wang et al., 2026a,d; Kang et al., 2025; Pang et al., 2026). Bounded cache management therefore defines the evolving belief state on which real-time reasoning depends.

4.3 Proactive and Interactive Video Assistants

Proactive assistants turn bounded state into interaction policies, deciding when evidence justifies an action, a clarification, or deliberate silence. Systems such as AURA and LiveStar unify continuous perception with multi-round dialogue under strict latency (Lu et al., 2026; Yang et al., 2026b; Tian et al., 2026; Wang et al., 2026b; Xiong et al., 2025), while MMDuet2 and STRIDE optimize the "when-to-speak" decision and treat silence as an active choice (Zheng et al., 2026; Wang et al., 2025g; Kim et al., 2026a). Architecturally, decoupling lightweight monitoring from asynchronous generation (*e.g.*, Dispider, ROMA) keeps perception responsive during synthesis (Qian et al., 2025), and for ultra-long streams selective dynamic memories and scene-graph state tracking enable anticipation and task-driven attention (Wang et al., 2026b; Xiong et al., 2025; Ma et al., 2026; Yan et al., 2026b; Yang et al., 2025c).

4.4 Real-Time Evaluation and Response Timing

Current benchmarks (Table 3) reach high scores but mainly probe short-window question answering. A complete protocol should also measure *when* and *how* a model acts, including response latency, premature response and missed-event rates, memory growth, long-horizon consistency, interruption handling, and proactive usefulness, to expose deployment trade-offs between latency, evidence coverage, and temporal consistency.

Takeaway. Streaming understanding shifts the objective from answering after observation to acting during observation; strong systems must jointly optimize causal memory, latency, and response timing.

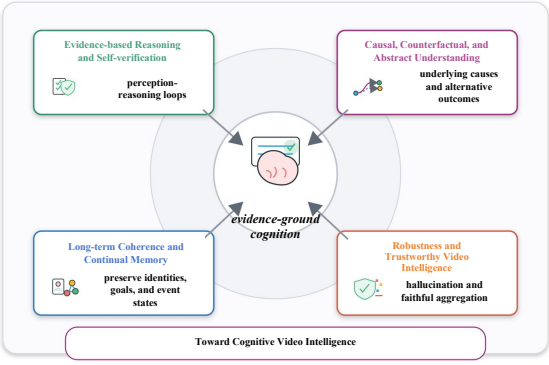

Figure 6: Toward cognitive video intelligence: open challenges.

5 Toward Cognitive Video Intelligence: Open Challenges

Long and streaming video understanding ultimately points beyond recognition toward cognitive video intelligence: models should reason over sparse evidence, maintain long-term consistency, explain causes, verify answers, and generalize reliably across dynamic environments.

Evidence-based reasoning and verification. Evidence sparsity makes reasoning inseparable from retrieval and grounding. Recent perception–reasoning loops use factuality-aware feedback or chain-of-thought over selected keyframes to support evidence-based answers (Pu et al., 2025; Sun et al., 2024b), and specialized domains such as tactical analysis, fine-grained sports, and mathematical reasoning further demand testable links among observation, inference, and justification (Chen et al., 2025c; He et al., 2025b). The open challenge is to move from answer-level supervision to process-level verification, where models expose evidence traces, cite temporal support, and indicate when the video is insufficient.

Long-term coherence and continual memory. Cognitive systems must preserve coherent state across long durations, repeated observations, and viewpoint changes. Inspired by hippocampal mechanisms, recent work studies continual learning and instance identity preservation to avoid forgetting and entity loss (Tan et al., 2026b; Liang et al., 2025b), while cross-view evaluations such as EgoExo-Con probe consistency between first- and third-person views (Jung et al., 2025). Memory thus becomes a commitment to preserve identities, goals, relations, and event states across offline long-video and always-on streaming assistants.

Causal, counterfactual, and abstract reasoning. A central challenge is distinguishing temporal association from causal explanation. Causal and counterfactual benchmarks ask models to explain anomalies and answer "what if" questions, requiring reasoning over underlying mechanisms and alternative outcomes (Yang et al., 2025a; Chen et al., 2025d; Zhu et al., 2025b; Cheng et al., 2026), a progression from co-occurrence to causal inference and finally to counterfactual simulation over physical, social, and procedural dynamics.

Robustness, generalization, and trustworthy intelligence. Reliability-oriented evaluations assess hallucination, omission, and faithful evidence aggregation (Rawal et al., 2025; Lu et al., 2025a; Yang et al., 2025b), while generalization studies test robustness under visual noise and cross-domain shifts such as underwater environments (Xue et al., 2026; Wang et al., 2025j). Benchmark integrity and evidence-grounded generation are thus complementary: uncontaminated evaluation reduces external leakage, while verifiable reasoning reduces unsupported internal claims.

Takeaway. Cognitive video intelligence should be evaluated as an evidence-grounded chain: perceive, remember, reason about causes and alternatives, and verify trustworthiness under distribution shifts.

6 Conclusion

This survey frames long and streaming video understanding as a shift from passive perception toward reliable temporal cognition. Benchmarks now probe long-context reasoning, multimodal evidence, cognitive abstraction, and benchmark integrity. Offline methods address the evidence bottleneck through input reduction, retrieval, memory, agentic search, and grounding, while streaming systems add causal processing, bounded state, response timing, and proactive interaction. Across both paradigms, the central design question is no longer how much video a model can ingest, but how reliably it can identify, organize, and verify task-relevant evidence under realistic constraints. Looking forward, progress will require co-designing evidence-grounded reasoning, long-horizon memory, and trustworthy evaluation, so that models not only answer correctly but also remain faithful, timely, and verifiable in dynamic open-world settings.

Limitations

This survey emphasizes representative benchmarks and methods that illustrate the transition toward reliable temporal cognition rather than exhaustively listing every concurrent system. Reported results are taken from original papers and are affected by heterogeneous protocols, sampling budgets, backbones, input lengths, and deployment settings, so our tables summarize trends rather than define strict leaderboards. The benchmark-overlap analysis is limited to publicly reported provenance and cannot rule out undisclosed training data or sample-level leakage. Finally, many cognitive and streaming capabilities still lack unified evaluation, especially for latency, efficiency, causal reasoning, and long-term memory.

References

Vatsal Agarwal, Saksham Suri, Matthew Gwilliam, Pulkit Kumar, and Abhinav Shrivastava. 2026. Going down memory lane: Scaling tokens for video stream understanding with dynamic kv-cache memory. *arXiv preprint arXiv:2602.18434*.

Anurag Arnab, Ahmet Iscen, Mathilde Caron, Alireza Fathi, and Cordelia Schmid. 2026. Temporal chain of thought: Long-video understanding by thinking in frames. *Advances in Neural Information Processing Systems*, 38:143018–143046.

Shehreen Azad, Vibhav Vineet, and Yogesh Singh Rawat. 2025. Hierarq: Task-aware hierarchical q-former for enhanced video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8545–8556.

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, and 1 others. 2025. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*.

Xiaoyi Bao, Chenwei Xie, Hao Tang, Tingyu Weng, Xiaofeng Wang, Yun Zheng, and Xingang Wang. 2025. Dynimg: Key frames with visual prompts are good representation for multi-modal video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23678–23688.

Yuxuan Cai, Jiangning Zhang, Zhenye Gan, Qingdong He, Xiaobin Hu, Junwei Zhu, Yabiao Wang, Chengjie Wang, Zhucun Xue, Chaoyou Fu, and 1 others. 2025. Humanvideo-mme: Benchmarking mllms for human-centric video understanding. *arXiv preprint arXiv:2507.04909*.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. 2024. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197.

Boyu Chen, Zikang Wang, Zhengrong Yue, Kainan Yan, Chenyun Yu, Yi Huang, Zijun Liu, Yafei Wen, Xiaoxin Chen, Yang Liu, and 1 others. 2025a. Videochat-m1: Collaborative policy planning for video understanding via multi-agent reinforcement learning. *arXiv preprint arXiv:2511.19524*.

Boyu Chen, Zhengrong Yue, Siran Chen, Zikang Wang, Yang Liu, Peng Li, and Yali Wang. 2025b. Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20237–20246.

Haodong Chen, Haojian Huang, Xinxiang Yin, and Dian Shao. 2025c. Finequest: Adaptive knowledge-assisted sports video understanding via agent-of-thoughts reasoning. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 2909–2918.

Houlun Chen, Xin Wang, Guangyao Li, Yuwei Zhou, Yihan Chen, Jia Jia, and Wenwu Zhu. 2026a. Think with grounding: Curriculum reinforced reasoning with video grounding for long video understanding. *arXiv preprint arXiv:2602.18702*.

Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang F Wang. 2024. Rextime: A benchmark suite for reasoning-across-time in videos. *Advances in Neural Information Processing Systems*, 37:28662–28673.

Kehua Chen. 2025. Chronoforge-rl: Chronological forging through reinforcement learning for enhanced video understanding. *arXiv preprint arXiv:2509.15800*.

Seng Nam Chen, Hao Chen, Chenglam Ho, Xinyu Mao, Jinping Wang, Yu Zhang, and Chao Li. 2026b. Seeing the scene matters: Revealing forgetting in video understanding models with a scene-aware long-video benchmark. *arXiv preprint arXiv:2603.27259*.

Wang Chen, Yongdong Luo, Yuhui Zeng, Luojun Lin, Tianyu Xie, Fei Chao, Rongrong Ji, and Xiawu Zheng. 2026c. Event-anchored frame selection for effective long-video understanding. *arXiv preprint arXiv:2603.00983*.

Wang Chen, Yuhui Zeng, Yongdong Luo, Tianyu Xie, Luojun Lin, Jiayi Ji, Yan Zhang, and Xiawu Zheng. 2026d. Wavelet-based frame selection by detecting semantic boundary for long video understanding. *arXiv preprint arXiv:2603.00512*.

Yuefei Chen, Jiang Liu, Xiaodong Lin, and Ruixiang Tang. 2025d. Countervqa: Evaluating and improving counterfactual reasoning in vision-language models for video understanding. *arXiv preprint arXiv:2511.19923*.

Yukang Chen, Wei Huang, Baifeng Shi, Qinghao Hu, Hanrong Ye, Ligeng Zhu, Zhijian Liu, Pavlo Molchanov, Jan Kautz, Xiaojuan Qi, and 1 others. 2026e. Scaling rl to long videos. *Advances in Neural Information Processing Systems*, 38:172842–172870.

Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, and 1 others. 2025e. Longvila: Scaling long-context visual language models for long videos. In *International Conference on Learning Representations*, volume 2025, pages 18227–18246.

Zhaoyu Chen, Hongnan Lin, Yongwei Nie, Fei Ma, Xuemiao Xu, Fei Yu, and Chengjiang Long. 2025f. Invert4tvg: A temporal video grounding framework with inversion tasks for enhanced action understanding. *arXiv preprint arXiv:2508.07388*.

Dingxin Cheng, Mingda Li, Jingyu Liu, Yongxin Guo, Bin Jiang, Qingbin Liu, Xi Chen, and Bo Zhao. 2025a. Enhancing long video understanding via hierarchical event-based memory. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

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

Ying Cheng, Yu-Ho Lin, Min-Hung Chen, Fu-En Yang, and Shang-Hong Lai. 2026. Vader: Towards causal video anomaly understanding with relation-aware large language models. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 7301–7311.

Janghoon Cho, Jungsoo Lee, Munawar Hayat, Kyuwoong Hwang, Fatih Porikli, and Sungha Choi. 2025. Floc: Facility location-based efficient visual token compression for long video understanding. *arXiv preprint arXiv:2511.00141*.

Meng Chu, Yicong Li, and Tat-Seng Chua. 2025. Graphvideoagent: Enhancing long-form video understanding with entity relation graphs. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4639–4648.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Ming Dai, Sen Yang, Boqiang Duan, Wankou Yang, and Jingdong Wang. 2025. Momentseg: Moment-centric sampling for enhanced video pixel understanding. *arXiv preprint arXiv:2510.09274*.

Shangzhe Di and Weidi Xie. 2024. Grounded question-answering in long egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12934–12943.

Shangzhe Di, Zhelun Yu, Guanghao Zhang, Haoyuan Li, Hao Cheng, Bolin Li, Wanggui He, Fangxun Shu, and Hao Jiang. 2025. Streaming video question-answering with in-context video kv-cache retrieval. In *International Conference on Learning Representations*, volume 2025, pages 42115–42127.

Anxhelo Diko, Tinghuai Wang, Wassim Swaileh, Shiyan Sun, and Ioannis Patras. 2025. Rewind: Understanding long videos with instructed learnable memory. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13734–13743.

Yang Ding, Yizhen Zhang, Xin Lai, Ruihang Chu, and Yujiu Yang. 2025. Videozoomer: Reinforcement-learned temporal focusing for long video reasoning. *arXiv preprint arXiv:2512.22315*.

Lars Doorenbos, Federico Spurio, and Juergen Gall. 2025. Video panels for long video understanding. *arXiv preprint arXiv:2509.23724*.

Bodong Du, Bowen Liu, Yang Yu, Xinpeng Ding, Zhiheng Wu, Shuning Wang, Shuo Nie, Naiming Liu, Qifeng Chen, Yangqiu Song, and 1 others. 2026. Medhorizon: Towards long-context medical video understanding in the wild. *arXiv preprint arXiv:2605.06537*.

Yifan Du, Kun Zhou, Yuqi Huo, Yifan Li, Wayne Xin Zhao, Haoyu Lu, Zijia Zhao, Bingning Wang, Weipeng Chen, and Ji-Rong Wen. 2024. Towards event-oriented long video understanding. *arXiv preprint arXiv:2406.14129*.

Sunqi Fan, Meng-Hao Guo, and Shuojin Yang. 2025a. Agentic keyframe search for video question answering. *arXiv preprint arXiv:2503.16032*.

Yue Fan, Xiaojian Ma, Rongpeng Su, Jun Guo, Rujie Wu, Xi Chen, and Qing Li. 2025b. Embodied videoagent: Persistent memory from egocentric videos and embodied sensors enables dynamic scene understanding. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 6342–6352. IEEE.

Bo Fang, Wenhao Wu, Qiangqiang Wu, Yuxin Song, and Antoni B Chan. 2025a. Threading keyframe with narratives: Mllms as strong long video comprehenders. *arXiv preprint arXiv:2505.24158*.

Pengcheng Fang, Yuxia Chen, and Rui Guo. 2025b. When and what: Diffusion-grounded videollm with entity aware segmentation for long video understanding. *arXiv preprint arXiv:2508.15641*.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124.

Gueter Josmy Faure, Jia-Fong Yeh, Min-Hung Chen, Hung-Ting Su, Shang-Hong Lai, and Winston H Hsu. 2025. Hermes: temporal-coherent long-form understanding with episodes and semantics. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22911–22921.

Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. 2024. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025a. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24108–24118.

Chaoyou Fu, Haozhi Yuan, Yuhao Dong, Yi-Fan Zhang, Yunhang Shen, Xiaoxing Hu, Xueying Li, Jinsen Su, Chengwu Long, Xiaoyao Xie, and 1 others. 2026. Video-mme-v2: Towards the next stage in benchmarks for comprehensive video understanding. *arXiv preprint arXiv:2604.05015*.

Shenghao Fu, Qize Yang, Yuan-Ming Li, Xihan Wei, Xiaohua Xie, and Wei-Shi Zheng. 2025b. Love-r1: Advancing long video understanding with an adaptive zoom-in mechanism via multi-step reasoning. *arXiv preprint arXiv:2509.24786*.

Hong Gao, Yiming Bao, Xuezhen Tu, Yutong Xu, Yue Jin, Yiyang Mu, Bin Zhong, Linan Yue, and Min-Ling Zhang. 2025. Agentic video intelligence: A flexible framework for advanced video exploration and understanding. *arXiv preprint arXiv:2511.14446*.

Shibo Gao, Peipei Yang, Yangyang Liu, Yi Chen, Han Zhu, Xu-Yao Zhang, and Linlin Huang. 2026a. Vagu & gts: Llm-based benchmark and framework for joint video anomaly grounding and understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 4167–4175.

Zhe Gao, Shiyu Shen, Taifeng Chai, Weinong Wang, Haotian Xu, Wenbin Li, Qi Fan, Yang Gao, Dacheng Tao, and 1 others. 2026b. Videotir: Accurate understanding for long videos with efficient tool-integrated reasoning. *arXiv preprint arXiv:2603.25021*.

Arushi Goel, Sreyan Ghosh, Vatsal Agarwal, Nishit Anand, Kaousheik Jayakumar, Lasha Koroshinadze, Yao Xu, Katie Lyons, James Case, Karan Sapra, and 1 others. 2026. Mmou: A massive multi-task omni understanding and reasoning benchmark for long and complex real-world videos. *arXiv preprint arXiv:2603.14145*.

Gabriele Goletto, Tushar Nagarajan, Giuseppe Averta, and Dima Damen. 2024. Amego: Active memory from long egocentric videos. In *European Conference on Computer Vision*, pages 92–110. Springer.

Kristen Grauman, Andrew Westbury, Lorenzo Torresani, Kris Kitani, Jitendra Malik, Triantafyllos Afouras, Kumar Ashutosh, Vijay Baiyya, Siddhant Bansal, Bikram Boote, and 1 others. 2024. Ego-exo4d: Understanding skilled human activity from first-and third-person perspectives. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19383–19400.

Xin Gu, Bing Fan, Jiali Yao, Zhipeng Zhang, Yan Huang, Cheng Han, Heng Fan, and Libo Zhang. 2026. Towards long-form spatio-temporal video grounding. *arXiv preprint arXiv:2602.23294*.

Dong Guo, Faming Wu, Feida Zhu, Fuxing Leng, Guang Shi, Haobin Chen, Haoqi Fan, Jian Wang, Jianyu Jiang, Jiawei Wang, and 1 others. 2025a. Seed1. 5-vl technical report. *arXiv preprint arXiv:2505.07062*.

Yanan Guo, Wenhui Dong, Jun Song, Shiding Zhu, Xuan Zhang, Hanqing Yang, Yingbo Wang, Yang Du, Xianing Chen, and Bo Zheng. 2025b. Fila-video: Spatio-temporal compression for fine-grained long video understanding. *arXiv preprint arXiv:2504.20384*.

Yongxin Guo, Jingyu Liu, Mingda Li, Dingxin Cheng, Xiaoying Tang, Dianbo Sui, Qingbin Liu, Xi Chen, and Kevin Zhao. 2025c. Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 3302–3310.

Yongxin Guo, Jingyu Liu, Mingda Li, Qingbin Liu, Xi Chen, and Xiaoying Tang. 2024. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*.

Zhenghui Guo, Yuanbin Man, Junyuan Sheng, Bowen Lin, Ahmed Ahmed, Bo Jiang, Boyuan Zhang, Miao Yin, Sian Jin, Omprakash Gnawal, and 1 others. 2026. Event-vstream: Event-driven real-time understanding for long video streams. *arXiv preprint arXiv:2601.15655*.

Zhuoning Guo, Mingxin Li, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Xiaowen Chu. 2025d. Towards universal video retrieval: Generalizing video embedding via synthesized multimodal pyramid curriculum. *arXiv preprint arXiv:2510.27571*.

ZhaoYang Han, Qihan Lin, Hao Liang, Bowen Chen, Zhou Liu, and Wentao Zhang. 2025. Longinsightbench: A comprehensive benchmark for evaluating omni-modal models on human-centric long-video understanding. *arXiv preprint arXiv:2510.17305*.

Zifan Han, Hongbo Sun, Jinglin Xu, Canhui Tang, Yulong Lei, Xuchong Zhang, Hongbin Sun, Zhongjiang He, and Hao Sun. 2026. Wat: Online video understanding needs watching before thinking. *arXiv preprint arXiv:2603.13412*.

Tanveer Hannan, Md Mohaiminul Islam, Jindong Gu, Thomas Seidl, and Gedas Bertasius. 2025. Revision-llm: Recursive vision-language model for temporal grounding in hour-long videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19012–19022.

Bo He, Hengduo Li, Young Kyun Jang, Menglin Jia, Xuefei Cao, Ashish Shah, Abhinav Shrivastava, and Ser-Nam Lim. 2024. Ma-lmm: Memory-augmented large multimodal model for long-term video understanding. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 13504–13514.

Jianxiang He, Meisheng Hong, Jungang Li, Weiyu Guo, Xuming Hu, and Hui Xiong. 2025a. Vsi: Visual subtitle integration for keyframe selection to enhance long video understanding. *arXiv preprint arXiv:2508.06869*.

Xusheng He, Wei Liu, Shanshan Ma, Qian Liu, Chenghao Ma, and Jianlong Wu. 2025b. Finebadminton: A multi-level dataset for fine-grained badminton video understanding. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12776–12783.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. 2025c. Framethinker: Learning to think with long videos via multi-turn frame spotlighting. *arXiv preprint arXiv:2509.24304*.

Zefeng He, Xiaoye Qu, Yafu Li, Siyuan Huang, Daizong Liu, and Yu Cheng. 2025d. Videossr: Video self-supervised reinforcement learning. *arXiv preprint arXiv:2511.06281*.

Jack Hong, Shilin Yan, Jiayin Cai, Xiaolong Jiang, Yao Hu, and Weidi Xie. 2025. Worldsense: Evaluating real-world omnimodal understanding for multimodal llms. *arXiv preprint arXiv:2502.04326*.

Jian Hu, Zixu Cheng, Chenyang Si, Wei Li, and Shaogang Gong. 2025a. Cos: Chain-of-shot prompting for long video understanding. *arXiv preprint arXiv:2502.06428*.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025b. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*.

Pengfei Hu, Meng Cao, Yingyao Wang, Yi Wang, Jiahua Dong, Jun Song, Yu Cheng, Bo Zheng, and Xiaodan Liang. 2025c. Thinking with drafts: Speculative temporal reasoning for efficient long video understanding. *arXiv preprint arXiv:2512.00805*.

Zi-Yuan Hu, Shuo Liang, Duo Zheng, Yanyang Li, Yeyao Tao, Shijia Huang, Wei Feng, Jia Qin, Jianguang Yu, Jing Huang, and 1 others. 2025d. Nemo: Needle in a montage for video-language understanding. *arXiv preprint arXiv:2509.24563*.

Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. 2024. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280.

De-An Huang, Subhashree Radhakrishnan, Zhiding Yu, and Jan Kautz. 2025a. Frag: Frame selection augmented generation for long video and long document understanding. *arXiv preprint arXiv:2504.17447*.

Yuning Huang, Xiaoyu Ji, Joseph Huang, Yichi Zhang, and Fengqing Zhu. 2026a. Adaptive greedy frame selection for long video understanding. *arXiv preprint arXiv:2603.20180*.

Yuxiang Huang, Mingye Li, Xu Han, Chaojun Xiao, Weilin Zhao, Ao Sun, Ziqi Yuan, Hao Zhou, Fandong Meng, and Zhiyuan Liu. 2026b. Spava: Accelerating long-video understanding via sequence-parallelism-aware approximate attention. *arXiv preprint arXiv:2601.21444*.

Zeyi Huang, Yuyang Ji, Xiaofang Wang, Nikhil Mehta, Tong Xiao, Donghyun Lee, Sigmund Vanvalkenburgh, Shengxin Zha, Bolin Lai, Licheng Yu, and 1 others. 2025b. Building a mind palace: Structuring environment-grounded semantic graphs for effective long video analysis with llms. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 24169–24179.

Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, and 1 others. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*.

Lawrence Jang, Yinheng Li, Dan Zhao, Charles Ding, Justin Lin, Paul Pu Liang, Rogerio Bonatti, and Kazuhito Koishida. 2025. Videowebarena: Evaluating long context multimodal agents with video understanding web tasks. In *International Conference on Learning Representations*, volume 2025, pages 36934–36958.

Jindong Jiang, Xiuyu Li, Zhijian Liu, Muyang Li, Guo Chen, Zhiqi Li, De-An Huang, Guilin Liu, Zhiding Yu, Kurt Keutzer, and 1 others. 2025a. Storm: Token-efficient long video understanding for multimodal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5830–5841.

Yifan Jiang, Yueying Wang, Rui Zhao, Toufiq Parag, Zhimin Chen, Zhenyu Liao, and Jayakrishnan Unnikrishnan. 2025b. Videop2r: Video understanding from perception to reasoning. *arXiv preprint arXiv:2511.11113*.

Hongbo Jin, Qingyuan Wang, Wenhao Zhang, Yang Liu, and Sijie Cheng. 2025. Videomem: Enhancing ultra-long video understanding via adaptive memory management. *arXiv preprint arXiv:2512.04540*.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual

representation empowers large language models with image and video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13700–13710.

Minjoon Jung, Junbin Xiao, Junghyun Kim, Byoung-Tak Zhang, and Angela Yao. 2025. Egoexo-con: Exploring view-invariant video temporal understanding. *arXiv preprint arXiv:2510.26113*.

Kumara Kahatapitiya, Kanchana Ranasinghe, Jongwoo Park, and Michael S Ryoo. 2025. Language repository for long video understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5627–5646.

Hyolim Kang, Yunsu Park, Youngbeom Yoo, Yeeun Choi, and Seon Joo Kim. 2025. Open-ended hierarchical streaming video understanding with vision language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20715–20725.

Geewook Kim and Minjoon Seo. 2025. Mambamia: A state-space-model-based compression for efficient video understanding in large multimodal models. *arXiv preprint arXiv:2506.13564*.

Junho Kim, Hosu Lee, James M Rehg, Minsu Kim, and Yong Man Ro. 2026a. Stride: When to speak meets sequence denoising for streaming video understanding. *arXiv preprint arXiv:2603.27593*.

Minsoo Kim, Kyuhong Shim, Jungwook Choi, and Simyung Chang. 2026b. Infinipot-v: Memory-constrained kv cache compression for streaming video understanding. *Advances in Neural Information Processing Systems*, 38:138983–139013.

Younggun Kim, Ahmed S Abdelrahman, and Mohamed Abdel-Aty. 2025. Vru-accident: A vision-language benchmark for video question answering and dense captioning for accident scene understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 761–771.

Fanqi Kong, Weiqin Zu, Xinyu Chen, Yaodong Yang, Song-Chun Zhu, and Xue Feng. 2025. Siv-bench: A video benchmark for social interaction understanding and reasoning. *arXiv preprint arXiv:2506.05425*.

Benno Krojer, Mojtaba Komeili, Candace Ross, Quentin Garrido, Koustuv Sinha, Nicolas Ballas, and Mahmoud Assran. 2025. A shortcut-aware video-qa benchmark for physical understanding via minimal video pairs. *arXiv preprint arXiv:2506.09987*.

Xiaohan Lan, Yitian Yuan, Zequn Jie, and Lin Ma. 2024. Vidcompress: Memory-enhanced temporal compression for video understanding in large language models. *arXiv preprint arXiv:2410.11417*.

Bo Li, Yuanhan Zhang, Liangyu Chen, Jinghao Wang, Fanyi Pu, Joshua Adrian Cahyono, Jingkang Yang, Chunyuan Li, and Ziwei Liu. 2025a. Otter: A multi-modal model with in-context instruction tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

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

Chenglin Li, Qianglong Chen, Yin Zhang, and 1 others. 2025b. Iterative zoom-in: Temporal interval exploration for long video understanding. *arXiv preprint arXiv:2507.02946*.

Handong Li, Zikang Liu, Longteng Guo, Tongtian Yue, Yepeng Tang, Xinxin Zhu, Chuanyang Zheng, Ziming Wang, Zhibin Wang, Jun Song, and 1 others. 2026a. Adaspark: Adaptive sparsity for efficient long-video understanding. *arXiv preprint arXiv:2604.08077*.

Hongyu Li, Jinyu Chen, Ziyu Wei, Shaofei Huang, Tianrui Hui, Jialin Gao, Xiaoming Wei, and Si Liu. 2025c. Llava-st: A multimodal large language model for fine-grained spatial-temporal understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8592–8603.

Hongyu Li, Songhao Han, Yue Liao, Junfeng Luo, Jialin Gao, Shuicheng Yan, and Si Liu. 2025d. Reinforcement learning tuning for videollms: Reward design and data efficiency. *arXiv preprint arXiv:2506.01908*.

Jiahua Li, Kun Wei, Zhe Xu, Zibo Su, Xu Yang, and Cheng Deng. 2025e. Perceive, reflect and understand long video: Progressive multi-granular clue exploration with interactive agents. *arXiv preprint arXiv:2509.24943*.

Jiaze Li, Hao Yin, Wenhui Tan, Jingyang Chen, Boshen Xu, Yuxun Qu, Yijing Chen, Jianzhong Ju, Zhenbo Luo, and Jian Luan. 2025f. Revisor: Beyond textual reflection, towards multimodal introspective reasoning in long-form video understanding. *arXiv preprint arXiv:2511.13026*.

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

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Ruanjun Li, Yuedong Tan, Yuanming Shi, and Jiawei Shao. 2025h. Videoscan: Enabling efficient streaming video understanding via frame-level semantic carriers. *arXiv preprint arXiv:2503.09387*.

Rui Li, Xiaohan Wang, Yuhui Zhang, Orr Zohar, Zeyu Wang, and Serena Yeung-Levy. 2025i. Temporal preference optimization for long-form video understanding. *arXiv preprint arXiv:2501.13919*.

Wei Li, Hehe Fan, Yongkang Wong, Mohan Kankanhalli, and Yi Yang. 2024c. Topa: Extending large language models for video understanding via text-only pre-alignment. *Advances in Neural Information Processing Systems*, 37:5697–5738.

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2024d. Llama-vid: An image is worth 2 tokens in large language models. In *European Conference on Computer Vision*, pages 323–340. Springer.

Yayuan Li, Aadit Jain, Filippos Bellos, and Jason J Corso. 2025j. Mistake attribution: Fine-grained mistake understanding in egocentric videos. *arXiv preprint arXiv:2511.20525*.

Zeqian Li, Shangzhe Di, Zhonghua Zhai, Weilin Huang, Yanfeng Wang, and Weidi Xie. 2026b. Universal video temporal grounding with generative multimodal large language models. *Advances in Neural Information Processing Systems*, 38:64426–64455.

Baoyu Liang, Qile Su, Shoutai Zhu, Yuchen Liang, and Chao Tong. 2025a. Videvent: A large dataset for understanding dynamic evolution of events in videos. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 5128–5136.

Hao Liang, Jiapeng Li, Tianyi Bai, Xijie Huang, Linzhuang Sun, Zhengren Wang, Conghui He, Bin Cui, Chong Chen, and Wentao Zhang. 2024. Keyvideollm: Towards large-scale video keyframe selection. *arXiv preprint arXiv:2407.03104*.

Yujia Liang, Jile Jiao, Xuetao Feng, Zixuan Ye, Yuan Wang, and Zhicheng Wang. 2025b. Ipformer-videollm: Enhancing multi-modal video understanding for multi-shot scenes. *arXiv preprint arXiv:2506.21116*.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26689–26699.

Jingyang Lin, Jialian Wu, Ximeng Sun, Ze Wang, Jiang Liu, Yusheng Su, Xiaodong Yu, Hao Chen, Jiebo Luo, Zicheng Liu, and 1 others. 2026. Unleashing hour-scale video training for long video-language understanding. *Advances in Neural Information Processing Systems*, 38:17523–17552.

Shih-Yao Lin, Sibendu Paul, and Caren Chen. 2025a. From captions to keyframes: Keyscore for multimodal frame scoring and video-language understanding. *arXiv preprint arXiv:2510.06509*.

Yueqian Lin, Jingyang Zhang, Qinsi Wang, Hancheng Ye, Yuzhe Fu, Yudong Liu, Hai Li, Yiran Chen, and 1 others. 2025b. Hippomm: Hippocampal-inspired multimodal memory for long audiovisual event understanding. *arXiv preprint arXiv:2504.10739*.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. 2025a. World model on million-length video and language with blockwise ringattention. In *International Conference on Learning Representations*, volume 2025, pages 45953–45977.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306.

Ruyang Liu, Chen Li, Yixiao Ge, Thomas H Li, Ying Shan, and Ge Li. 2024b. Bt-adapter: Video conversation is feasible without video instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13658–13667.

Ruyang Liu, Chen Li, Haoran Tang, Yixiao Ge, Ying Shan, and Ge Li. 2024c. St-llm: Large language models are effective temporal learners. In *European Conference on Computer Vision*, pages 1–18. Springer.

Shuming Liu, Chen Zhao, Tianqi Xu, and Bernard Ghanem. 2025b. Bolt: Boost large vision-language model without training for long-form video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 3318–3327.

Xiangrui Liu, Yan Shu, Zheng Liu, Ao Li, Yang Tian, and Bo Zhao. 2025c. Video-xl-pro: Reconstructive token compression for extremely long video understanding. *arXiv preprint arXiv:2503.18478*.

Ye Liu, Kevin Qinghong Lin, Chang Wen Chen, and Mike Zheng Shou. 2025d. Videomind: A chain-of-lora agent for long video reasoning. *arXiv e-prints*, pages arXiv–2503.

Xinwei Long, Kai Tian, Peng Xu, Guoli Jia, Jingxuan Li, Sa Yang, Yihua Shao, Kaiyan Zhang, Che Jiang, Hao Xu, and 1 others. 2025. Adsqa: Towards advertisement video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 23396–23407.

Hao Lu, Jiahao Wang, Yaolun Zhang, Ruohui Wang, Xuanyu Zheng, Yepeng Tang, Dahua Lin, and Lewei Lu. 2025a. Elv-halluc: Benchmarking semantic aggregation hallucinations in long video understanding. *arXiv preprint arXiv:2508.21496*.

Haoyu Lu, Yuqi Huo, Yifan Du, Zijia Zhao, Longteng Guo, Bingning Wang, Weipeng Chen, and Jing Liu. 2025b. Needle in a video haystack: A scalable synthetic evaluator for video mllms. In *International Conference on Learning Representations*, volume 2025, pages 99750–99782.

Xudong Lu, Yang Bo, Jinpeng Chen, Shuhan Li, Xintong Guo, Huankang Guan, Fang Liu, Dunyuan Xu, Peiwen Sun, Heyang Sun, and 1 others. 2026. Aura:

Always-on understanding and real-time assistance via video streams. *arXiv preprint arXiv:2604.04184*.

Fuwen Luo, Shengfeng Lou, Chi Chen, Ziyue Wang, Chenliang Li, Weizhou Shen, Jiyue Guo, Peng Li, Ming Yan, Ji Zhang, and 1 others. 2025. Museg: Reinforcing video temporal understanding via timestamp-aware multi-segment grounding. *arXiv preprint arXiv:2505.20715*.

David Ma, Huaqing Yuan, Xingjian Wang, Qianbo Zang, Tianci Liu, Xinyang He, Yanbin Wei, Jiawei Guo, Ni Jiahui, Zhenzhu Yang, and 1 others. 2025a. Scalelong: A multi-timescale benchmark for long video understanding. *arXiv preprint arXiv:2505.23922*.

Ke Ma, Jiaqi Tang, Bin Guo, Xueting Han, Ruonan Xu, Qingfeng He, Ziheng Wang, Xu Wang, Qifeng Chen, Zhiwen Yu, and 1 others. 2026. Response-g1: Explicit scene graph modeling for proactive streaming video understanding. *arXiv preprint arXiv:2605.07575*.

Wentao Ma, Weiming Ren, Yiming Jia, Zhuofeng Li, Ping Nie, Ge Zhang, and Wenhu Chen. 2025b. Videoeval-pro: Robust and realistic long video understanding evaluation. *arXiv preprint arXiv:2505.14640*.

Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. 2024. 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 (Volume 1: Long Papers)*, pages 12585–12602.

Neelu Madan, Andreas Møgelmose, Rajat Modi, Yogesh S Rawat, and Thomas B Moeslund. 2024. Foundation models for video understanding: A survey. *arXiv preprint arXiv:2405.03770*.

Yuanbin Man, Ying Huang, Chengming Zhang, Bingzhe Li, Wei Niu, and Miao Yin. 2025. Adacmˆ 2: On understanding extremely long-term video with adaptive cross-modality memory reduction. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8534–8544.

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

Jiahao Meng, Tan Yue, Qi Xu, Haochen Wang, Zhongwei Ren, Weisong Liu, Yuhao Wang, Renrui Zhang, Yunhai Tong, and Haodong Duan. 2026. Videozerobench: Probing the limits of video mllms with spatio-temporal evidence verification. *arXiv preprint arXiv:2604.01569*.

Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. 2025. Videoglamm: A large multimodal model for pixel-level visual grounding in videos. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19036–19046.

Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See Kiong Ng, and Luu Anh Tuan. 2024. Video-language understanding: A survey from model architecture, model training, and data perspectives. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3636–3657.

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

Zhanzhong Pang, Dibyadip Chatterjee, Fadime Sener, and Angela Yao. 2026. Decouple and cache: Kv cache construction for streaming video understanding. *arXiv preprint arXiv:2605.01858*.

Ziqi Pang and Yu-Xiong Wang. 2025. Mr. video:" mapreduce" is the principle for long video understanding. *arXiv preprint arXiv:2504.16082*.

Shrenik Patel and Daivik Patel. 2025. Cacheflow: Compressive streaming memory for efficient long-form video understanding. *arXiv preprint arXiv:2511.13644*.

Taiying Peng, Jiacheng Hua, Miao Liu, and Feng Lu. 2026a. In the eye of mllm: Benchmarking egocentric video intent understanding with gaze-guided prompting. *Advances in Neural Information Processing Systems*, 38.

Tianhao Peng, Haochen Wang, Yuanxing Zhang, Noah Wang, Zili Wang, Ge Zhang, Jian Yang, Shihao Li, Yanghai Wang, Xintao Wang, and 1 others. 2026b. Mvu-eval: Towards multi-video understanding evaluation for multimodal llms. *Advances in Neural Information Processing Systems*, 38.

Bowei Pu, Chuanbin Liu, Yifan Ge, Peicheng Zhou, Yiwei Sun, Zhiying Lu, Jiankang Wang, and Hongtao Xie. 2025. Alternating perception-reasoning for hallucination-resistant video understanding. *arXiv preprint arXiv:2511.18463*.

Haozhe Qi, Kevin Qu, Mahdi Rad, Rui Wang, Alexander Mathis, and Marc Pollefeys. 2026. Adapttoken: Entropy-based adaptive token selection for mllm long video understanding. *arXiv preprint arXiv:2603.28696*.

Rui Qian, Shuangrui Ding, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Yuhang Cao, Dahua Lin, and Jiaqi Wang. 2025. 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.

Jihao Qiu, Lingxi Xie, Xinyue Huo, Qi Tian, and Qixiang Ye. 2026a. Longvideo-r1: Smart navigation for low-cost long video understanding. *arXiv preprint arXiv:2602.20913*.

Zongyang Qiu, Bingyuan Wang, Xingbei Chen, Yingqing He, and Zeyu Wang. 2026b. Emovid: A multimodal emotion video dataset for emotion-centric video understanding and generation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 8612–8620.

Ali Rasekh, Erfan Soula, Omid Daliran, Simon Gottschalk, Mohsen Fayyaz, and 1 others. 2026. Enhancing temporal understanding in video-llms through stacked temporal attention in vision encoders. *Advances in Neural Information Processing Systems*, 38:83112–83139.

Ruchit Rawal, Reza Shirkavand, Heng Huang, Gowthami Somepalli, and Tom Goldstein. 2025. Argus: Hallucination and omission evaluation in video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20280–20290.

Aniket Rege, Arka Sadhu, Yuliang Li, Kejie Li, Ramya Korlakai Vinayak, Yuning Chai, Yong Jae Lee, and Hyo Jin Kim. 2026. Agentic very long video understanding. *arXiv preprint arXiv:2601.18157*.

Shuhuai Ren, Sishuo Chen, Shicheng Li, Xu Sun, and Lu Hou. 2023. Testa: Temporal-spatial token aggregation for long-form video-language understanding. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 932–947.

Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. 2024. 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.

Weiming Ren, Wentao Ma, Huan Yang, Cong Wei, Ge Zhang, and Wenhu Chen. 2025. Vamba: Understanding hour-long videos with hybrid mamba-transformers. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21197–21208.

Vishnu Sai, Dheeraj Sai, Girish Varma, Priyesh Shukla, and 1 others. 2026. Dual-signal adaptive kv-cache optimization for long-form video understanding in vision-language models. *arXiv preprint arXiv:2602.14236*.

Saul Santos, António Farinhas, Daniel C McNamee, and Andre Martins. 2025. ∞-video: A training-free approach to long video understanding via continuous-time memory consolidation. In *International Conference on Machine Learning*, pages 52877–52893. PMLR.

Kuleen Sasse, Efsun Sarioglu Kayi, and Arun Reddy. 2025. Controllable hybrid captioner for improved long-form video understanding. *arXiv preprint arXiv:2507.17047*.

Ashish Seth, Utkarsh Tyagi, Ramaneswaran Selvakumar, Nishit Anand, Sonal Kumar, Sreyan Ghosh, Ramani Duraiswami, Chirag Agarwal, and Dinesh Manocha. 2025. Egoillusion: Benchmarking hallucinations in egocentric video understanding. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 28449–28468.

Sahil Shah, SP Sharan, Harsh Goel, Minkyu Choi, Mustafa Munir, Manvik Pasula, Radu Marculescu, and Sandeep Chinchali. 2026. Neus-qa: Grounding long-form video understanding in temporal logic and neuro-symbolic reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 8805–8813.

Xiaoqian Shen, Min-Hung Chen, Yu-Chiang Frank Wang, Mohamed Elhoseiny, and Ryo Hachiuma. 2025. Zoom-zero: Reinforced coarse-to-fine video understanding via temporal zoom-in. *arXiv preprint arXiv:2512.14273*.

Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, and 1 others. 2024. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint arXiv:2410.17434*.

Yuan Sheng, Yanbin Hao, Chenxu Li, Shuo Wang, and Xiangnan He. 2025. Sevices: Unifying semantic-visual evidence consensus for long video understanding. *arXiv preprint arXiv:2510.20622*.

Zhengpeng Shi, Hengli Li, Yanpeng Zhao, Jianqun Zhou, Yuxuan Wang, Qinrong Cui, Wei Bi, Songchun Zhu, Bo Zhao, and Zilong Zheng. 2025. V-hub: A visual-centric humor understanding benchmark for video llms. *arXiv preprint arXiv:2509.25773*.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, and 1 others. 2024. 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.

Enxin Song, Wenhao Chai, Shusheng Yang, Ethan Armand, Xiaojun Shan, Haiyang Xu, Jianwen Xie, and Zhuowen Tu. 2025a. Videonsa: Native sparse attention scales video understanding. *arXiv preprint arXiv:2510.02295*.

Enxin Song, Wenhao Chai, Tian Ye, Jenq-Neng Hwang, Xi Li, and Gaoang Wang. 2025b. Moviechat+: Question-aware sparse memory for long video question answering. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Boyuan Sun, Jiaxing Zhao, Xihan Wei, and Qibin Hou. 2025a. Llava-scissor: Token compression with semantic connected components for video llms. *arXiv preprint arXiv:2506.21862*.

Guangyu Sun, Archit Singhal, Burak Uzkent, Mubarak Shah, Chen Chen, and Garin Kessler. 2025b. From frames to clips: Training-free adaptive key clip selection for long-form video understanding. *arXiv preprint arXiv:2510.02262*.

Shangkun Sun, Ruyang Liu, Haoran Tang, Yixiao Ge, Haibo Lu, Wei Gao, Jiankun Yang, and Chen Li. 2024a. Ppllava: Varied video sequence understanding with prompt guidance. *arXiv preprint arXiv:2411.02327*.

Yiwei Sun, Zhihang Liu, Chuanbin Liu, Bowei Pu, Zhihan Zhang, and Hongtao Xie. 2024b. Hallucination mitigation prompts long-term video understanding. *arXiv preprint arXiv:2406.11333*.

Wenhui Tan, Xiaoyi Yu, Jiaze Li, Yijing Chen, Jianzhong Ju, Zhenbo Luo, Ruihua Song, and Jian Luan. 2026a. Msjoe: Jointly evolving mllm and sampler for efficient long-form video understanding. *arXiv preprint arXiv:2602.22932*.

Xichen Tan, Yuanjing Luo, Yunfan Ye, Fang Liu, and Zhiping Cai. 2025. Allvb: All-in-one long video understanding benchmark. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 7211–7219.

Yue Tan, Xiaoqian Hu, Hao Xue, Celso De Melo, and Flora Salim. 2026b. Bisecle: Binding and separation in continual learning for video language understanding. *Advances in Neural Information Processing Systems*, 38:33752–33782.

Canhui Tang, Zifan Han, Hongbo Sun, Sanping Zhou, Xuchong Zhang, Xin Wei, Ye Yuan, Huayu Zhang, Jinglin Xu, and Hao Sun. 2026a. Tspo: Temporal sampling policy optimization for long-form video language understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 9368–9376.

Qiance Tang, Ziqi Wang, Jieyu Lin, Ziyun Li, Barbara De Salvo, and Sai Qian Zhang. 2026b. Egoeverything: A benchmark for human behavior inspired long context egocentric video understanding in ar environment. *arXiv preprint arXiv:2604.08342*.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, and 1 others. 2025. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*.

Keda Tao, Can Qin, Haoxuan You, Yang Sui, and Huan Wang. 2025. Dycoke: Dynamic compression of tokens for fast video large language models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18992–19001.

Keda Tao, Yuhua Zheng, Jia Xu, Wenjie Du, Kele Shao, Hesong Wang, Xueyi Chen, Xin Jin, Junhan Zhu, Bohan Yu, and 1 others. 2026. Lvomnibench: Pioneering long audio-video understanding evaluation for omnimodal llms. *arXiv preprint arXiv:2603.19217*.

Shulin Tian, Ruiqi Wang, Hongming Guo, Penghao Wu, Yuhao Dong, Xiuying Wang, Jingkang Yang, Hao Zhang, Hongyuan Zhu, and Ziwei Liu. 2025. Ego-r1: Chain-of-tool-thought for ultra-long egocentric video reasoning. *arXiv preprint arXiv:2506.13654*.

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

Ashwin Vinod, Shrey Pandit, Aditya Vavre, and Linshen Liu. 2025. Egovlm: Policy optimization for egocentric video understanding. *arXiv preprint arXiv:2506.03097*.

Chao Wang, Xudong Tan, Jianjian Cao, Kangcong Li, and Tao Chen. 2026a. Curvestream: Boosting streaming video understanding in mllms via curvature-aware hierarchical visual memory management. *arXiv preprint arXiv:2603.19571*.

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

Haibo Wang, Chenghang Lai, Yixuan Sun, and Weifeng Ge. 2024a. Weakly supervised gaussian contrastive grounding with large multimodal models for video question answering. In *Proceedings of the 32nd ACM International Conference on Multimedia*, pages 5289–5298.

Haibo Wang, Zhiyang Xu, Yu Cheng, Shizhe Diao, Yufan Zhou, Yixin Cao, Qifan Wang, Weifeng Ge, and Lifu Huang. 2024b. Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. *arXiv preprint arXiv:2410.03290*.

Hao Wang, Eiki Murata, Lingfang Zhang, Ayako Sato, So Fukuda, Ziqi Yin, Wentao Hu, Keisuke Nakao, Yusuke Nakamura, Sebastian Zwirner, and 1 others. 2026c. Vir-bench: Evaluating geospatial and temporal understanding of mllms via travel video itinerary reconstruction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 9747–9756.

Junxi Wang, Te Sun, Jiayi Zhu, Junxian Li, Haowen Xu, Zichen Wen, Xuming Hu, Zhiyu Li, and Linfeng Zhang. 2026d. Streammeco: Long-term agent memory compression for efficient streaming video understanding. *arXiv preprint arXiv:2604.09000*.

Lei Wang, Piotr Koniusz, and Yongsheng Gao. 2025a. Video understanding by design: How datasets

shape architectures and insights. *arXiv preprint arXiv:2509.09151*.

Maijunxian Wang, Ruisi Wang, Juyi Lin, Ran Ji, Thaddäus Wiedemer, Qingying Gao, Dezhi Luo, Yaoyao Qian, Lianyu Huang, Zelong Hong, and 1 others. 2026e. A very big video reasoning suite. *arXiv preprint arXiv:2602.20159*.

Shaobo Wang, Tianle Niu, Runkang Yang, Deshan Liu, Xu He, Zichen Wen, Conghui He, Xuming Hu, and Linfeng Zhang. 2025b. Videocompressa: Data-efficient video understanding via joint temporal compression and spatial reconstruction. *arXiv preprint arXiv:2511.18831*.

Shaoguang Wang, Weiyu Guo, Ziyang Chen, Xuming Hu, and Hui Xiong. 2026f. Where to focus: Query-modulated multimodal keyframe selection for long video understanding. *arXiv preprint arXiv:2604.17422*.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, and 1 others. 2025c. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967.

Xiao Wang, Qingyi Si, Jianlong Wu, Shiyu Zhu, Li Cao, and Liqiang Nie. 2024c. Retake: Reducing temporal and knowledge redundancy for long video understanding. *arXiv preprint arXiv:2412.20504*.

Xiao Wang, Qingyi Si, Shiyu Zhu, Jianlong Wu, Li Cao, and Liqiang Nie. 2025d. Adaretake: Adaptive redundancy reduction to perceive longer for video-language understanding. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5417–5432.

Xiaoqi Wang, Yi Wang, and Lap-Pui Chau. 2025e. Eva02-at: Egocentric video-language understanding with spatial-temporal rotary positional embeddings and symmetric optimization. *arXiv preprint arXiv:2506.14356*.

Ye Wang, Ziheng Wang, Boshen Xu, Yang Du, Kejun Lin, Zihan Xiao, Zihao Yue, Jianzhong Ju, Liang Zhang, Dingyi Yang, and 1 others. 2026g. Time-r1: Post-training large vision language model for temporal video grounding. *Advances in Neural Information Processing Systems*, 38:83330–83364.

Yi Wang, Xinhao Li, Ziang Yan, Yinan He, Jiashuo Yu, Xiangyu Zeng, Chenting Wang, Changlian Ma, Haian Huang, Jianfei Gao, and 1 others. 2025f. Internvideo2.5: Empowering video mllms with long and rich context modeling. *arXiv preprint arXiv:2501.12386*.

Youze Wang, Zijun Chen, Ruoyu Chen, Shishen Gu, Wenbo Hu, Jiayang Liu, Yinpeng Dong, Hang Su, Jun Zhu, Meng Wang, and 1 others. 2026h. Benchmarking trustworthiness in multimodal llms for video understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 37979–37987.

Yueqian Wang, Songxiang Liu, Disong Wang, Nuo Xu, Guanglu Wan, Huishuai Zhang, and Dongyan Zhao. 2025g. Mmduet2: Enhancing proactive interaction of video mllms with multi-turn reinforcement learning. *arXiv preprint arXiv:2512.06810*.

Zheng Wang, Haoran Chen, Haoxuan Qin, Zhipeng Wei, Tianwen Qian, and Cong Bai. 2026i. Think, then verify: A hypothesis-verification multi-agent framework for long video understanding. *arXiv preprint arXiv:2603.04977*.

Zikang Wang, Boyu Chen, Zhengrong Yue, Yi Wang, Yu Qiao, Limin Wang, and Yali Wang. 2026j. Videochat-a1: Thinking with long videos by chain-of-shot reasoning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 10467–10475.

Ziyang Wang, Shoubin Yu, Elias Stengel-Eskin, Jaehong Yoon, Feng Cheng, Gedas Bertasius, and Mohit Bansal. 2025h. 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.

Ziyang Wang, Yue Zhang, Shoubin Yu, Ce Zhang, Zengqi Zhao, Jaehong Yoon, Hyunji Lee, Gedas Bertasius, and Mohit Bansal. 2026k. Egomemreason: A memory-driven reasoning benchmark for long-horizon egocentric video understanding. *arXiv preprint arXiv:2605.09874*.

Ziyang Wang, Honglu Zhou, Shijie Wang, Junnan Li, Caiming Xiong, Silvio Savarese, Mohit Bansal, Michael S Ryoo, and Juan Carlos Niebles. 2025i. Active video perception: Iterative evidence seeking for agentic long video understanding. *arXiv preprint arXiv:2512.05774*.

Ziyi Wang, Zhi Gao, Boxuan Yu, Zirui Dai, Yuxiang Song, Qingyuan Lu, Jin Chen, and Xinxiao Wu. 2025j. Vudg: A dataset for video understanding domain generalization. *arXiv preprint arXiv:2505.24346*.

Hang Wu, Sherin Mary Mathews, Yujun Cai, Ming-Hsuan Yang, and Yiwei Wang. 2026a. Semantic-aware adaptive visual memory for streaming video understanding. *arXiv preprint arXiv:2605.07897*.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. 2024. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857.

Tz-Ying Wu, Sharath Nittur Sridhar, and Subarna Tripathi. 2026b. Harnessing object grounding for time-sensitive video understanding. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 2475–2484.

Zhirong Wu, Xiaodong Wang, Langling Huang, Teng Xu, and Peixi Peng. 2026c. A training-free framework for long video understanding via video-query-options similarity. In *The Fourteenth International Conference on Learning Representations*.

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

Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. 2024. Funqa: Towards surprising video comprehension. In *European Conference on Computer Vision*, pages 39–57. Springer.

Yuan Xie, Tianshui Chen, Zheng Ge, and Lionel Ni. 2025. Video-mtr: Reinforced multi-turn reasoning for long video understanding. *arXiv preprint arXiv:2508.20478*.

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

Boshen Xu, Zihan Xiao, Jiaze Li, Jianzhong Ju, Zhenbo Luo, Jian Luan, and Qin Jin. 2025a. Timeviper: A hybrid mamba-transformer vision-language model for efficient long video understanding. *arXiv preprint arXiv:2511.16595*.

Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. 2024. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*.

Weili Xu, Enxin Song, Wenhao Chai, Xuexiang Wen, Tian Ye, and Gaoang Wang. 2025b. Auroralong: Bringing rnns back to efficient open-ended video understanding. *arXiv preprint arXiv:2507.02591*.

Xinyi Xu, Hongsong Wang, Guo-Sen Xie, Caifeng Shan, and Fang Zhao. Hitea: Hierarchical temporal alignment for training-free long-video temporal grounding. In *The Fourteenth International Conference on Learning Representations*.

Yicheng Xu, Yue Wu, Jiashuo Yu, Ziang Yan, Tianxiang Jiang, Yinan He, Qingsong Zhao, Kai Chen, Yu Qiao, Limin Wang, and 1 others. 2025c. Expvid: A benchmark for experiment video understanding & reasoning. *arXiv preprint arXiv:2510.11606*.

Ziqiang Xu, Qi Dai, Tian Xie, Yifan Yang, Kai Qiu, DongDong Chen, Zuxuan Wu, and Chong Luo. 2025d. Viarl: Adaptive temporal grounding via visual iterated amplification reinforcement learning. *arXiv preprint arXiv:2505.15447*.

Xizhe Xue, Yang Zhou, Dawei Yan, Lijie Tao, Junjie Li, Ying Li, Haokui Zhang, and Rong Xiao. 2026. Uvlm: Benchmarking video language model for underwater world understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 11532–11540.

Cui Yakun, Peng Qi, Fushuo Huo, Hang Du, Weijie Shi, Juntao Dai, Zhenghao Zhu, Sirui Han, and Yike Guo. 2025. Perception, understanding and reasoning, a multimodal benchmark for video fake news detection. *arXiv preprint arXiv:2510.24816*.

Sosuke Yamao, Natsuki Miyahara, Yuankai Qi, and Shun Takeuchi. 2026. Question-guided visual compression with memory feedback for long-term video understanding. *arXiv preprint arXiv:2603.15167*.

Haiyang Yan, Hongyun Zhou, Peng Xu, Xiaoxue Feng, and Mengyi Liu. 2026a. Symphony: A cognitively-inspired multi-agent system for long-video understanding. *arXiv preprint arXiv:2603.17307*.

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

Ziang Yan, Yinan He, Xinhao Li, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. 2026c. Videochat-r1. 5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *Advances in Neural Information Processing Systems*, 38:119152–119184.

Cheng Yang, Haiyuan Wan, Yiran Peng, Xin Cheng, Zhaoyang Yu, Jiayi Zhang, Junchi Yu, Xinlei Yu, Xiawu Zheng, Dongzhan Zhou, and 1 others. 2025a. Reasoning via video: The first evaluation of video models' reasoning abilities through maze-solving tasks. *arXiv preprint arXiv:2511.15065*.

Garry Yang, Zizhe Chen, Man Hon Wong, Haoyu Lei, Yongqiang Chen, Zhenguo Li, Kaiwen Zhou, and James Cheng. 2025b. Mesh-understanding videos like human: Measuring hallucinations in large video models. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 4827–4836.

Haolin Yang, Feilong Tang, Lingxiao Zhao, Xinlin Zhuang, Yifan Lu, Xiang An, Ming Hu, Xiaofeng Zhang, Abdalla Swikir, Junjun He, and 1 others. 2025c. Streamagent: Towards anticipatory agents for streaming video understanding. *arXiv preprint arXiv:2508.01875*.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, and 1 others. 2025d. Egolife: Towards egocentric life assistant. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28885–28900.

Ruoliu Yang, Chu Wu, Caifeng Shan, Ran He, and Chaoyou Fu. 2026a. Videodetective: Clue hunting via both extrinsic query and intrinsic relevance for long video understanding. *arXiv preprint arXiv:2603.22285*.

Te Yang, Xiangyu Zhu, Bo Wang, Quan Chen, Peng Jiang, and Zhen Lei. 2025e. Eea: Exploration-exploitation agent for long video understanding. *arXiv preprint arXiv:2512.03500*.

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

Zeyuan Yang, Delin Chen, Xueyang Yu, Maohao Shen, and Chuang Gan. 2025g. Vca: Video curious agent for long video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20168–20179.

Zhenyu Yang, Kairui Zhang, Yuhang Hu, Bing Wang, Shengsheng Qian, Bin Wen, Fan Yang, Tingting Gao, Weiming Dong, and Changsheng Xu. 2026b. Livestar: Live streaming assistant for real-world online video understanding. *Advances in Neural Information Processing Systems*, 38:31266–31304.

Zhongyu Yang, Zuhao Yang, Shuo Zhan, Tan Yue, Wei Pang, and Yingfang Yuan. 2026c. Svagent: Storyline-guided long video understanding via cross-modal multi-agent collaboration. *arXiv preprint arXiv:2604.05079*.

Zhoufaran Yang, Yan Shu, Jing Wang, Zhifei Yang, Yan Zhang, Yu Li, Keyang Lu, Gangyan Zeng, Shaohui Liu, Yu Zhou, and 1 others. 2025h. Vidtext: Towards comprehensive evaluation for video text understanding. *arXiv preprint arXiv:2505.22810*.

Yifeng Yao, Yike Yun, Jing Wang, Huishuai Zhang, Dongyan Zhao, Ke Tian, Zhihao Wang, Minghui Qiu, and Tao Wang. 2025. K-frames: Scene-driven any-k keyframe selection for long video understanding. *arXiv preprint arXiv:2510.13891*.

Hanrong Ye, Haotian Zhang, Erik Daxberger, Lin Chen, Zongyu Lin, Yanghao Li, Bowen Zhang, Haoxuan You, Dan Xu, Zhe Gan, and 1 others. 2025a. Mmego: Towards building egocentric multimodal llms for video qa. In *International Conference on Learning Representations*, volume 2025, pages 71705–71723.

Jinhui Ye, Zihan Wang, Haosen Sun, Keshigeyan Chandrasegaran, Zane Durante, Cristobal Eyzaguirre, Yonatan Bisk, Juan Carlos Niebles, Ehsan Adeli, Li Fei-Fei, and 1 others. 2025b. Re-thinking temporal search for long-form video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8579–8591.

Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, and 1 others. 2023. mplug-owl: Modularization empowers large language models with multimodality. *arXiv preprint arXiv:2304.14178*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, and Fei Huang. 2024. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. In *Proceedings of the ieee/cvf conference on computer vision and pattern recognition*, pages 13040–13051.

Jeong Hun Yeo, Sangyun Chung, Sungjune Park, Dae Hoe Kim, Jinyoung Moon, and Yong Man Ro. 2025. Gcagent: Long-video understanding via schematic and narrative episodic memory. *arXiv preprint arXiv:2511.12027*.

Yufei Yin, Yuchen Xing, Qianke Meng, Minghao Chen, Yan Yang, and Zhou Yu. 2026. Progressive video condensation with mllm agent for long-form video understanding. *arXiv preprint arXiv:2604.02891*.

Zeng You, Zhiquan Wen, Yaofo Chen, Xin Li, Runhao Zeng, Yaowei Wang, and Mingkui Tan. 2024. Toward long video understanding via fine-detailed video story generation. *IEEE Transactions on Circuits and Systems for Video Technology*, 35(5):4592–4607.

An Yu, Weiheng Lu, Jian Li, Zhenfei Zhang, Yunhang Shen, Felix X-F Ye, and Ming-Ching Chang. 2025a. Smart: Shot-aware multimodal video moment retrieval with audio-enhanced mllm. *arXiv preprint arXiv:2511.14143*.

Jiashuo Yu, Yue Wu, Meng Chu, Zhifei Ren, Zizheng Huang, Pei Chu, Ruijie Zhang, Yinan He, Qirui Li, Songze Li, and 1 others. 2025b. Vrbench: A benchmark for multi-step reasoning in long narrative videos. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 21655–21666.

Xueqing Yu, Bohan Li, Yan Li, and Zhenheng Yang. 2026a. Virtuebench: Evaluating trustworthiness under uncertainty in long video understanding. *arXiv preprint arXiv:2603.07071*.

Yating Yu, Congqi Cao, Zhaoying Wang, Weihua Meng, Jie Li, Yuxin Li, Zihao Wei, Zhongpei Shen, and Jiajun Zhang. 2026b. Cuebench: Advancing unified understanding of context-aware video anomalies in real-world. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 12187–12195.

Chao Yuan, Yang Yang, Yehui Yang, and Zach Cheng. 2025a. Date: Dynamic absolute time enhancement for long video understanding. *arXiv preprint arXiv:2509.09263*.

Huaying Yuan, Zheng Liu, Junjie Zhou, Ji-Rong Wen, and Zhicheng Dou. 2025b. Videodeepresearch: Long video understanding with agentic tool using. *arXiv e-prints*, pages arXiv–2506.

Huaying Yuan, Jian Ni, Zheng Liu, Yueze Wang, Junjie Zhou, Zhengyang Liang, Bo Zhao, Zhao Cao, Ji-Rong Wen, and Zhicheng Dou. 2026. Momentseeker: A task-oriented benchmark for long-video moment retrieval. *Advances in Neural Information Processing Systems*, 38.

Xiangyu Zeng, Kunchang Li, Chenting Wang, Xinhao Li, Tianxiang Jiang, Ziang Yan, Songze Li, Yansong Shi, Zhengrong Yue, Yi Wang, and 1 others. 2025. Timesuite: Improving mllms for long video understanding via grounded tuning. In *International Conference on Learning Representations*, volume 2025, pages 38057–38081.

Ce Zhang, Taixi Lu, Md Mohaiminul Islam, Ziyang Wang, Shoubin Yu, Mohit Bansal, and Gedas Bertasius. 2024a. A simple llm framework for long-range video question-answering. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21715–21737.

Dell Zhang, Xiangyu Chen, Jixiang Luo, Mengxi Jia, Changzhi Sun, Ruilong Ren, Jingren Liu, Hao Sun, and Xuelong Li. 2025a. Infinite video understanding. *arXiv preprint arXiv:2507.09068*.

Hang Zhang, Xin Li, and Lidong Bing. 2023. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the 2023 conference on empirical methods in natural language processing: system demonstrations*, pages 543–553.

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

Haowei Zhang, Shudong Yang, Jinlan Fu, See-Kiong Ng, and Xipeng Qiu. 2026a. Hermes: Kv cache as hierarchical memory for efficient streaming video understanding. *arXiv preprint arXiv:2601.14724*.

Hongzhi Zhang, Jingyuan Zhang, Xingguang Ji, Qi Wang, and Fuzheng Zhang. 2025c. Dyntok: Dynamic compression of visual tokens for efficient and effective video understanding. *arXiv preprint arXiv:2506.03990*.

Huaxin Zhang, Xiaohao Xu, Xiang Wang, Jialong Zuo, Xiaonan Huang, Changxin Gao, Shanjun Zhang, Li Yu, and Nong Sang. 2025d. Holmes-vau: Towards long-term video anomaly understanding at any granularity. In *Proceedings of the computer vision and pattern recognition conference*, pages 13843–13853.

Kairui Zhang, Zhenyu Yang, Bing Wang, Shengsheng Qian, and Changsheng Xu. 2026b. Querystream: Advancing streaming video understanding with query-aware pruning and proactive response. In *The Fourteenth International Conference on Learning Representations*.

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

Shaojie Zhang, Jiahui Yang, Jianqin Yin, Zhenbo Luo, and Jian Luan. 2025e. Q-frame: Query-aware frame selection and multi-resolution adaptation for video-llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22056–22065.

Shaolei Zhang, Qingkai Fang, Yang Yang, and Yang Feng. 2025f. Llava-mini: Efficient image and video large multimodal models with one vision token. In *International Conference on Learning Representations*, volume 2025, pages 53285–53310.

Shuyi Zhang, Xiaoshuai Hao, Yingbo Tang, Lingfeng Zhang, Pengwei Wang, Zhongyuan Wang, Hongxuan Ma, and Shanghang Zhang. 2025g. Video-cot: A comprehensive dataset for spatiotemporal understanding of videos based on chain-of-thought. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 12745–12752.

Xiaoyi Zhang, Zhaoyang Jia, Zongyu Guo, Jiahao Li, Bin Li, Houqiang Li, and Yan Lu. 2026c. Deep video discovery: Agentic search with tool use for long-form video understanding. *Advances in Neural Information Processing Systems*, 38:89863–89895.

Yiming Zhang, Zhuokai Zhao, Zhaorun Chen, Zenghui Ding, Xianjun Yang, and Yining Sun. 2025h. Beyond training: Dynamic token merging for zero-shot video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22046–22055.

Yuanhan Zhang, Yunice Chew, Yuhao Dong, Aria Leo, Bo Hu, and Ziwei Liu. 2025i. Towards video thinking test: A holistic benchmark for advanced video reasoning and understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 20626–20636.

Yulin Zhang, Cheng Shi, and Sibei Yang. 2026d. Weavetime: Stream from earlier frames into emergent memory in videollms. *arXiv preprint arXiv:2602.22142*.

Zhen Zhang, Yuhang Yang, Yunxiang Jiang, Yuhuan Lu, Haifeng Lu, Zheng Lian, Runhao Zeng, and Xiping Hu. 2026e. Affectseek: Agentic affective understanding in long videos under vague user queries. *arXiv preprint arXiv:2605.05640*.

Zhihong Zhang, Xiaojian Huang, Jin Xu, Zhuodong Luo, Xinzhi Wang, Jiansheng Wei, and Xuejin Chen. 2025j. Videorewardbench: Comprehensive evaluation of multimodal reward models for video understanding. *arXiv preprint arXiv:2509.00484*.

Tiancheng Zhao, Qianqian Zhang, Kyusong Lee, Peng Liu, Lu Zhang, Chunxin Fang, Jiajia Liao, Kelei Jiang, Yibo Ma, and Ruochen Xu. 2024. Omchat: A recipe to train multimodal language models with strong long context and video understanding. *arXiv preprint arXiv:2407.04923*.

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

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

Yangyang Zhong, Ji Qi, Yuan Yao, Pengxin Luo, Yunfeng Yan, Donglian Qi, Zhiyuan Liu, and Tat-Seng Chua. 2025. Textvidbench: A benchmark for long video scene text understanding. *arXiv preprint arXiv:2506.04983*.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2(5):6.

Wenqi Zhou, Kai Cao, Hao Zheng, Yunze Liu, Xinyi Zheng, Miao Liu, Per Ola Kristensson, Walterio Mayol-Cuevas, Fan Zhang, Weizhe Lin, and 1 others. 2025a. X-lebench: A benchmark for extremely long egocentric video understanding. *arXiv preprint arXiv:2501.06835*.

Yiyang Zhou, Yangfan He, Yaofeng Su, Siwei Han, Joel Jang, Gedas Bertasius, Mohit Bansal, and Huaxiu Yao. 2026a. Reagent-v: A reward-driven multi-agent framework for video understanding. *Advances in Neural Information Processing Systems*, 38:151454–151491.

Yuan Zhou, Litao Hua, Shilong Jin, Wentao Huang, and Haoran Duan. 2026b. Reason: Reinforced causal search with information bottleneck for video understanding. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pages 13791–13799.

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

Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, Shenglong Ye, Lixin Gu, Hao Tian, Yuchen Duan, Weijie Su, Jie Shao, and 1 others. 2025a. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*.

Liyun Zhu, Qixiang Chen, Xi Shen, and Xiaodong Cun. 2025b. Vau-r1: Advancing video anomaly understanding via reinforcement fine-tuning. *arXiv preprint arXiv:2505.23504*.

Nannan Zhu, Yonghao Dong, Teng Wang, Xueqian Li, Shengjun Deng, Yijia Wang, Zheng Hong, Tiantian Geng, Guo Niu, Hanyan Huang, and 1 others. 2025c. Cvbench: Evaluating cross-video synergies for complex multimodal understanding and reasoning. *arXiv e-prints*, pages arXiv–2508.

Yuehan Zhu, Jingqi Zhao, Jiawen Zhao, Xudong Mao, and Baoquan Zhao. 2026. Hicrew: Hierarchical reasoning for long-form video understanding via question-aware multi-agent collaboration. *arXiv preprint arXiv:2604.21444*.

Orr Zohar, Xiaohan Wang, Yann Dubois, Nikhil Mehta, Tong Xiao, Philippe Hansen-Estruch, Licheng Yu, Xiaofang Wang, Felix Juefei-Xu, Ning Zhang, and 1 others. 2025. Apollo: An exploration of video understanding in large multimodal models. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 18891–18901.

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

Heqing Zou, Tianze Luo, Guiyang Xie, Victor Xiao Jie Zhang, Fengmao Lv, Guangcong Wang, Junyang Chen, Zhuochen Wang, Hansheng Zhang, and Huaijian Zhang. 2025a. Hlv-1k: A large-scale hour-long video benchmark for time-specific long video understanding. In *2025 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6. IEEE.

Yuanhao Zou, Shengji Jin, Andong Deng, Youpeng Zhao, Jun Wang, and Chen Chen. 2025b. Air: Enabling adaptive, iterative, and reasoning-based frame selection for video question answering. *arXiv preprint arXiv:2510.04428*.

Jialong Zuo, Yongtai Deng, Lingdong Kong, Jingkang Yang, Rui Jin, Yiwei Zhang, Nong Sang, Liang Pan, Ziwei Liu, and Changxin Gao. 2026. Videolucy: Deep memory backtracking for long video understanding. *Advances in Neural Information Processing Systems*, 38:25660–25691.

A Full Long-Video Benchmark Results

Table 4 provides the complete long-video benchmark result table corresponding to the condensed body Table 2. It is placed in the appendix because the main text keeps only representative cited methods to stay within the page budget.

B Full Streaming Benchmark Results

Table 5 provides the complete streaming benchmark table corresponding to the condensed body Table 3.

C Benchmark Provenance Lists

This appendix separates evaluation benchmarks by whether our manual literature-based provenance review identified potentially overlapping training datasets. Benchmarks without identified overlap are listed in Table 6; benchmarks with identified overlap are listed in Table 7. For each benchmark, we inspect the reported data sources in benchmark papers and compare them with the training-data sources reported by Video-LLM papers.

C.1 Benchmarks without Identified Overlap

Table 6 reports the complete set of benchmarks for which our provenance review did not identify any overlapping training dataset. The rows are grouped by broad evaluation scope only for readability.

C.2 Potentially Contaminated Benchmarks with Identified Overlap

Table 7 lists benchmarks for which at least one potentially overlapping training dataset was identified. Results on these benchmarks should therefore be interpreted cautiously when evaluating models trained on public video-data mixtures.

Methods	Baseline	F/S	Video-MME (11–3600s)	LongVideoBench (15–3600s)	LVBench (∼4101s)	MLVU (180–7200s)
Proprietary Model						
GPT-4o	-	-	71.9	66.7	34.7	64.6
GPT-4.1	-	-	72.0	-	63.4	-
Gemini-1.5-Flash	-	-	70.3	61.6	-	-
Gemini-1.5-Pro	-	-	75.0	64.0	33.1	-
Gemini-2.5-Flash	-	-	74.2	66.2	56.7	72.4
Gemini-2.5-Pro	-	-	82.4	69.8	67.4	79.6
Base Video-LLMs and MLLMs						
Seed1.5-VL	-	-	77.9	74.4	64.6	82.1
Qwen2.5-VL-7B	-	-	65.1	60.7	45.3	68.3
Qwen3-VL-8B	-	-	79.2	-	67.7	84.3
VAMBA-10B	-	-	57.8	55.9	42.1	65.9
Apollo-7B	-	-	61.3	58.5	-	70.9
VideoChat-R1.5-7B-S	-	-	65.2	61.4	46.3	70.1
VideoChat-R1.5-7B-M	-	-	67.1	62.6	48.4	70.9
Token Compression						
ReTaKe	LLaVA-Video-7B	-	64.0	59.7	48.5	67.8
AdaRETAKE	LLaVA-Video-7B	-	64.0	59.6	49.6	70.6
Flow4Agent	LLaVA-Video-7B	64	64.7	60.4	-	71.4
PPLLaVA	LLaVA-Video-7B	-	64.5	60.4	-	-
ReTaKe	Qwen2-VL-7B	-	63.9	57.7	47.8	69.8
AdaRETAKE	Qwen2-VL-7B	-	64.2	57.2	48.9	72.0
AdaptToken	Qwen2.5-VL	-	70.5	65.2	54.8	76.8
AdaRETAKE	Qwen2.5-VL-7B	-	67.7	62.6	51.2	75.0
AdaSpark	Qwen2.5-VL-7B	-	66.2	62.1	47.9	69.8
Keyframe Selection						
FTT	LLaVA-Video-7B	32	66.52	61.41	46.16	73.38
TSPO	LLaVA-Video-7B	64	65.5	63.9	45.3	76.3
EFS	LLaVA-Video-7B	64	65.6	62.1	-	70.9
SeViCES	LLaVA-Video-7B	64	65.6	63.1	47.3	73.1
Nar-KFC	LLaVA-Video-7B	61.6	61.6	-	67.7	-
ZoomV	LLaVA-Video-7B	80	64.0	60.9	50.0	68.1
EFS	LLaVA-OneVision-7B	8	57.4	60.3	-	67.4
FRAG	LLaVA-OneVision-7B	-	58.6	57.6	-	65.3
Nar-KFC	LLaVA-OneVision-7B	57.8	56.5	-	66.2	-
FTT	InternVL2.5-8B	32	65.37	60.66	46.61	69.98
SeViCES	InternVL2.5-8B	64	64.7	61.7	46.7	72.1
ZoomV	InternVL2.5-8B	80	64.4	63.3	51.5	70.0
Q-Frame	Qwen2-VL-7B	44	58.3	58.4	-	65.4
Nar-KFC	Qwen2-VL-7B	56.7	53.6	-	68.5	-
LOVE-R1	Qwen2.5-VL-7B	800	66.2	60.1	48.2	67.4
TSPO	Qwen2.5-VL-7B	64	65.5	64.2	46.4	74.3
DATE	Qwen2.5-VL-7B	256	67.3	63.3	47.4	-
EFS	Qwen2.5-VL-7B	16	60.0	60.5	-	66.0
F2C	Qwen2.5-VL-7B	8	61.4	58.9	-	63.1
F2C	Qwen2.5-VL-7B	16	63.4	61.1	-	65.5
F2C	Qwen2.5-VL-7B	32	65.6	60.8	-	66.8
K-frames	Qwen2.5-VL-7B	8	57.4	-	57.7	60.4
K-frames	Qwen2.5-VL-7B	32	62.1	-	60.5	65.9
K-frames	Qwen2.5-VL-7B	64	64.5	-	61.6	67.8
MSJoE	Qwen2.5-VL-7B	32	64.3	60.1	46.4	69.3
MSJoE	Qwen2.5-VL-7B	64	66.2	62.2	51.1	75.1
SeViCES	Qwen2.5-VL-7B	64	65.5	63.9	45.4	72.2
SpecTemp	Qwen2.5-VL-7B	14.5	62.4	57.5	-	48.6
SpecTemp	Qwen2.5-VL-7B	58.1	64.1	61.4	-	50.9
ZoomV	Qwen2.5-VL-7B	80	63.6	61.0	51.3	67.0
Memory-based Methods						
FlexMem	LLaVA-Video-7B	512/1024	64.7	63.6	51.0	72.4
FlexMem	LLaVA-OneVision-7B	512/1024	59.0	59.4	46.2	68.9
VideoDetective	Qwen2.5-VL-7B	-	62.1	-	51.1	45.8
Video-EM	Qwen2.5-VL-7B	27	62.0	-	45.7	-
VideoMem	Qwen3-VL-8B	-	73.6	63.3	58.5	77.4
DVD	API	-	67.3	71.6	74.2	-
HAVEN	API	-	-	78.2	81.0	-
Agentic Reasoning Methods						
VideoMind	Qwen2-VL-7B	-	58.2	-	40.8	64.4
ReAgent-V	Qwen2-VL-7B	33	58.3	46.4	35.6	56.1
SVAgent	Qwen2.5-VL-7B	8	61.2	60.7	40.6	62.7
VideoChat-A1	Qwen2.5-VL-7B	42.0	71.8	64.2	-	71.9
VideoDeepResearch	Qwen2.5-VL-7B	32	-	64.1	50.7	55.9
FrameThinker	Qwen2.5-VL-7B	-	-	52.9	36.6	59.1
ReAgent-V	Qwen2.5-VL-7B	35	60.7	54.3	40.5	60.7
REVISOR	Qwen2.5-VL-7B	-	65.7	57.5	42	69.8
Video-MTR	Qwen2.5-VL-7B	64	62.2	-	41.8	49.8
VideoZoomer	Qwen2.5-VL-7B	-	65.2	57.7	41.5	55.8
SVAgent	Qwen3-VL-8B	8	63.1	61.0	40.8	65.6
VideoChat-A1	InternVL2.5-8B	35.9	72.3	65.2	-	75.1
VideoChat-A1	InternVideo2.5-8B	41.3	72.9	65.4	-	76.2
VideoDeepResearch	Seed1.5-VL-pro	32	-	70.6	55.5	64.5
ReAgent-V	LLaVA-Video-7B	34	57.9	53.1	35.2	58.8
Symphony	DeepSeek-R1/DeepSeek-V3/Seed1.6-VL	-	-	77.1	71.8	81.0
AVP	Gemini-2.5-Flash	-	81.2	70.2	63.8	74.1
AVP	Gemini-2.5-Pro	-	85.3	73.4	74.8	84.3
MR.Video	Gemini-2.0-Flash/GPT-4o	-	61.8	61.6	60.8	-
Temporal Grounding						
Video-TwG(LR)	Qwen2.5-VL-7B	-	53.6	49.7	-	54.6
Video-TwG(HR)	Qwen2.5-VL-7B	-	59.7	56.3	-	60.3
ViTL	Qwen2.5-VL-7B	128	-	63.3	47.4	-
TPO	LLaVA-Video-7B	-	65.6	60.1	-	71.1

Table 4: Complete reported results on long-video understanding benchmarks. F/S denotes frames per sample. The grouped rows summarize proprietary and base Video-LLMs, compression, keyframe selection, memory, agentic reasoning, and temporal grounding methods. Because protocols, sampling budgets, backbones, and benchmark provenance differ, the table should be interpreted as a trend summary rather than a strict fair ranking; MLVU is also listed as potentially overlapped in Appendix Table 7.

Methods	Baseline	Frame/Sample	OVO-Bench	StreamingBench
Proprietary Model				
Gemini-1.5-Pro	-	1fps	70.8	75.69
GPT-4o	-	64	63.6	73.28
Base Video-LLMs				
Qwen2.5-VL-7B	-	2	75.0	76.39
Qwen2.5-VL-7B	-	4	78.4	78.47
Qwen2.5-VL-7B	-	8	76.6	79.11
Qwen3-VL-8B	-	2	79.3	78.31
Qwen3-VL-8B	-	4	81.4	80.59
Qwen3-VL-8B	-	8	79.9	78.83
WAT	Qwen2.5VL-7B	1fps	64.7	77.70
WeaveTime	LLaVA-OV-7B	1fps	68.82	72.13
WeaveTime	Qwen2-VL-7B	1fps	66.28	75.39
Token Compression				
FluxMem	Qwen2.5-VL-7B	1fps	63.0	76.4
QueryStream-7B	Qwen2.5-VL-7B	1fps	63.1	75.32
StreamingAssistant	Qwen2.5-VL-7B	-	56.1	69.16
Memory-based Methods				
EventMemAgent	Qwen3-VL-8B	≤ 32	68.29	77.0
FreshMem	Qwen2-VL-7B	-	66.67	74.20
SAVEMem	Qwen2.5-VL-7B	1fps	74.93	76.0
Streaming Cache				
CurveStream	LLaVA-OneVision-7B	10-20	70.57	75.12
DSCache	LLaVA-OneVision-7B	1fps	71.5	79.12
HERMES	LLaVA-OneVision-7B	0.5fps	66.34	73.23
CurveStream	Qwen2-VL-7B	10-20	70.73	81.04
CurveStream	Qwen2.5-VL-7B	10-20	73.48	84.00
DSCache	Qwen2.5-VL-7B	1fps	71.7	82.32
HERMES	Qwen2.5-VL-7B	1fps	68.98	79.44
CurveStream	Qwen3-VL-8B	10-20	80.76	85.56
Proactive Streaming Assistant				
StreamBridge	LLaVA-OneVision-7B	1fps	69.93	70.92
StreamBridge	Qwen2-VL-7B	1fps	71.30	77.04
Response-G1	Qwen3-VL-8B	-	73.6	77.5
AURA	Qwen3-VL-8B	-	79.8	83.2

Table 5: Full reported results on streaming video understanding benchmarks. The Frame/Sample column reports sampled frames per input or frame rate when marked as fps.

Evaluation Scope	Benchmarks without Identified Overlap
General video sources and captioning	HDVILA100M; MSRVTT; MSVD; MPII-MD; M-VAD; YFCC100M
Action and procedure understanding	Breakfast; CrossTask; Diving48; GTEA Gaze+; THUMOS14; THUMOS15; UCF101; YouTube Highlights
Long-video and chapter-level understanding	FineVideo; LVU; VidChapters-7M
Embodied, 3D, and scene understanding	HM3D; ScanNet; SceneWalK
Tracking and low-level dynamics	TrackingNet
Cognitive and reasoning-oriented evaluation	CG-Bench; CinePile; Video-Holmes; VLEP

Table 6: Full list of evaluation benchmarks for which no potentially overlapping training datasets were identified through manual provenance review.

Evaluation Benchmark	Potentially Overlapping Training Datasets
ActionBench	Ego4D, SSv2, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Grounded-VideoLLM, Haystack-Ego4D, InternVideo, LLaVA-OV, LLaVA-Video-178k, LongVid, MVBench, SpecTemp-80K, Tarsier2-Recap-585K, VideoChat2-IT, VideoExplorer, VideoMarathon, VM22M
ActivityNet	ActivityNet-Captions, ActivityNet-QA, DPO, E.T.Instruct 164K, GCG, LLaVA-OV, LLaVA-Video-178k, LongVR, ShareGPTVideo, SpecTemp-80K, Tarsier2-Recap-585K, TextVR, TimePro, VideoChatGPT, VideoExplorer, VideoMarathon
ActivityNet-Captions	ActivityNet, ActivityNet-QA, ActivityNet-RTL, DeVE-QA, DPO, E.T.Instruct 164K, GCG, LLaVA-OV, LLaVA-Video-178k, LongVR, ShareGPTVideo, SpecTemp-80K, Tarsier2-Recap-585K, TempGCap, TextVR, TimeIT, TimePro, VideoChatGPT, VideoExplorer, VideoMarathon, VideoXum, VTG-IT-120K, VTimeLLM
ActivityNet-CON	ActivityNet-Captions, ActivityNet, ActivityNet-RTL, DeVE-QA, TempGCap, TimeIT, TimePro, VideoXum, VTG-IT-120K, VTimeLLM
ActivityNet-QA	ActivityNet, ActivityNet-Captions, DPO, E.T.Instruct 164K, GCG, LLaVA-OV, LLaVA-Video-178k, LongVR, MMAT, ShareGPTVideo, SpecTemp-80K, Tarsier2-Recap-585K, TextVR, TimePro, VideoChatGPT, VideoExplorer, VideoMarathon
ActivityNet-RTL	ActivityNet Captions, ActivityNet, DeVE-QA, TempGCap, TimeIT, TimePro, VideoXum, VTG-IT-120K, VTimeLLM
AGQA	Charades, Charades-Ego, Charades-STA, LLaVA-OV, LLaVA-Video-178k, Tarsier2-Recap-585K
AMB	EPIC-KITCHENS, EgoIT-99K
AoTBench	NExT-QA, PerceptionTest, TemporalBench, TVBench, UCF101, VITATECS, Grounded-VideoLLM, Intent-QA, MVBench, M4-Instruct, SpecTemp-80K, VideoChat2-IT, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k
AVA	HCSTVG, InternVideo
BDD100K	Roadtext-1k
Causal-VidQA	Kinetics, FineAction, Grounded-VideoLLM, InternVideo, LLaVA-Video-178k, ShareGemini, Tarsier2-Recap-585K, VideoChat2-IT, VM22M
Charades-CON	Charades-STA, Charades, MVBench, TimeIT
Charades-Ego	Charades, EgoIT-99K, Tarsier2-Recap-585K
Charades-STA	Charades, Charades-Ego, LLaVA-OV, LLaVA-Video-178k, MVBench, Tarsier2-Recap-585K, TimeIT
CLEVRER	COF-DATA, Grounded-VideoLLM, MVBench, SpecTemp-80K, VideoChat2-IT, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k
COIN	E.T.Instruct 164K, Grounded-VideoLLM, TimeIT, TimePro, VTG-IT-120K
CUVA	COF-DATA, VideoEspresso
Daily-Omni	FineVideo
DiDeMo	YFCC100M, E.T.Instruct 164K, Grounded-VideoLLM, TempGCap, TimeIT, TimePro, TimeRFT, VTG-IT-120K, VTimeLLM
EgoCVR	Ego4D, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Haystack-Ego4D, LLaVA-OV, LLaVA-Video-178k, LongVid, SpecTemp-80K, Tarsier2-Recap-585K, VideoExplorer, VideoMarathon
EgoGazeVQA	Ego4D, EGTEA Gaze+, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Haystack-Ego4D, LLaVA-OV, LLaVA-Video-178k, LongVid, SpecTemp-80K, Tarsier2-Recap-585K, VideoExplorer, VideoMarathon
EgoPlan	EPIC-KITCHENS, Ego4D, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Haystack-Ego4D, LLaVA-OV, LLaVA-Video-178k, LongVid, SpecTemp-80K, Tarsier2-Recap-585K, VideoExplorer, VideoMarathon
EgoSchema	Ego4D, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Haystack-Ego4D, LLaVA-OV, LLaVA-Video-178k, LongVid, SpecTemp-80K, Tarsier2-Recap-585K, VideoExplorer, VideoMarathon
EGOTIMEQA	NaQ, EgoIT-99K, E.T.Instruct 164K
Ego4D	EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Haystack-Ego4D, LLaVA-OV, LLaVA-Video-178k, LongVid, SpecTemp-80K, Tarsier2-Recap-585K, VideoExplorer, VideoMarathon
EGTEA Gaze+	GTEA Gaze+, EgoIT-99K
Epic-Kitchens	EgoIT-99K
E.T. Bench	Charades-STA, CrossTask, Ego4D, HiREST, HT-Step, PerceptionTest, QAEgo4D, QVHighlights, STAR, SumMe, THUMOS14, THUMOS15, TVSum, YouCook2, YouTube Highlights, Charades, COF-DATA, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Grounded-VideoLLM, Haystack-Ego4D, HowTo100M, LLaVA-OV, LLaVA-Video-178k, LongVid, LongVR, MVBench, M4-Instruct, SpecTemp-80K, Tarsier2-Recap-585K, TextVR, TimeIT, TimePro, TimeRFT, VideoChat2-IT, VideoExplorer, VideoMarathon, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k, VTG-IT-120K, VideoEspresso
Event-Bench	EgoSchema, FunQA, NExT-QA, STAR, Grounded-VideoLLM, Intent-QA, MVBench, M4-Instruct, SpecTemp-80K, TimePro, VideoChat2-IT, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k
FCVID	FineAction

Table 7: Potentially contaminated evaluation benchmarks with identified training-data overlap (continued).

Evaluation Benchmark	Potentially Overlapping Training Datasets
FunQA	MVBench
GoT-10k	VideoChat-R1-18k
Haystack-Ego4D	Ego4D, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, LLaVA-OV, LLaVA-Video-178k, LongVid, SpecTemp-80K, Tarsier2-Recap-585K, VideoExplorer, VideoMarathon
HCSTVG-v1	AVA, HCSTVG, InternVideo
HiREST	HowTo100M, Grounded-VideoLLM, HowToCaption, HowToStep, HTStep, InternVideo, LongVid, TimeIT, TimePro, TimeRFT, VM22M
HourEval	MLVU
HourVideo	Ego4D, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Haystack-Ego4D, LLaVA-OV, LLaVA-Video-178k, LongVid, SpecTemp-80K, Tarsier2-Recap-585K, VideoExplorer, VideoMarathon
How2QA	HowTo100M, HowToCaption, HowToStep, HTStep, InternVideo, LongVid, TimeRFT, VM22M
HowTo100M	HowToCaption, HowToStep, HTStep, InternVideo, LongVid, TimeRFT, VM22M
HT-Step	HowTo100M, HowToCaption, HowToStep, InternVideo, LongVid, TimeRFT, VM22M
IntentQA	NExT-QA, Grounded-VideoLLM, M4-Instruct, VideoChat2-IT, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k
Kinetics	Causal-VidQA, FineAction, Grounded-VideoLLM, InternVideo, LLaVA-Video-178k, ShareGemini, Tarsier2-Recap-585K, VideoChat2-IT, VM22M
LongInsightBench	FineVideo
LongVideo-Reason	Shot2Story
LongViTU	Ego4D, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Haystack-Ego4D, LLaVA-OV, LLaVA-Video-178k, LongVid, SpecTemp-80K, Tarsier2-Recap-585K, VideoExplorer, VideoMarathon
LSDBench	Ego4D, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Haystack-Ego4D, LLaVA-OV, LLaVA-Video-178k, LongVid, SpecTemp-80K, Tarsier2-Recap-585K, VideoExplorer, VideoMarathon
LSMDC	MPII-MD, M-VAD, Tarsier2-Recap-585K
LV-HAYSTACK	Ego4D, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Haystack-Ego4D, LLaVA-OV, LLaVA-Video-178k, LongVid, SpecTemp-80K, Tarsier2-Recap-585K, VideoExplorer, VideoMarathon
MAD	LSMDC, Tarsier2-Recap-585K
MiraData	LongVid
MLVU	Ego4D, MovieChat-1K, COF-DATA, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Haystack-Ego4D, LLaVA-OV, LLaVA-Video-178k, LongVid, LongVR, SpecTemp-80K, Tarsier2-Recap-585K, TimePro, VideoExplorer, VideoMarathon, VideoEspresso
Moments in Time V1	MVBench
MovieChat-1K	COF-DATA, LongVR, SpecTemp-80K, TimePro, VideoMarathon, VideoEspresso
MovieNet	LLaMA-VID
MoVQA	MVBench
MSRVTT-QA	MSRVTT, MMAT
MSVC	MSRVTT, MSVD, VATEX, COF-DATA, Tarsier2-Recap-585K, VideoEspresso
MSVD-QA	MSVD, MMAT
MVBench	ActionBench, Charades-STA, CLEVRER, FunQA, Moments in Time V1, MoVQA, NTU RGB+D, PerceptionTest, STAR, TVQA, VLN-CE, Charades, COF-DATA, Grounded-VideoLLM, M4-Instruct, SpecTemp-80K, TimeIT, TimePro, VideoChat2-IT, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k
NaQ	E.T.Instruct 164K
NExT-GQA	NExT-QA, Grounded-VideoLLM, Intent-QA, M4-Instruct, SpecTemp-80K, VideoChat-R1-18k, VideoChat2-IT, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k
NExT-QA	VidOR, Grounded-VideoLLM, Intent-QA, LLaVA-Video-178k, M4-Instruct, VideoChat2-IT, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k
NTU RGB+D	MVBench
Oops	Tarsier2-Recap-585K
OpenEQA	HM3D, ScanNet
OpenVid-1M	LongVR
Panda-70M	Grounded-VideoLLM, Koala-36M, VideoExplorer, VideoMarathon
PerceptionTest	MVBench, SpecTemp-80K, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k
PVSG	VidOR, LLaVA-Video-178k
QAEgo4D	Ego4D, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Haystack-Ego4D, LLaVA-OV, LLaVA-Video-178k, LongVid, SpecTemp-80K, Tarsier2-Recap-585K, VideoExplorer, VideoMarathon
QuerYD	E.T.Instruct 164K, Grounded-VideoLLM, TimeIT, TimePro, TimeRFT, VTG-IT-120K
QVHighlights	TimeIT
ReXTime	ActivityNet, QVHighlights, ActivityNet-Captions, ActivityNet-QA, DPO, E.T.Instruct 164K, GCG, LLaVA-OV, LLaVA-Video-178k, LongVR, ShareGPTVideo, SpecTemp-80K, Tarsier2-Recap-585K, TextVR, TimeIT, TimePro, VideoChatGPT, VideoExplorer, VideoMarathon
RVS-Ego	Ego4D, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Haystack-Ego4D, LLaVA-OV, LLaVA-Video-178k, LongVid, SpecTemp-80K, Tarsier2-Recap-585K, VideoExplorer, VideoMarathon
RVS-Movie	MovieNet, LLaMA-VID
ShareGPT4Video	BDD100K, Ego4D, Panda-70M, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Grounded-VideoLLM, Haystack-Ego4D, Koala-36M, LLaVA-OV, LLaVA-Video-178k, LongVid, Roadtext-1k, SpecTemp-80K, Tarsier2-Recap-585K, VideoExplorer, VideoMarathon
Shot2Story	HDVILA100M, LongVideo-Reason
SQA3D	ScanNet, COF-DATA, VideoEspresso

Table 7: Potentially contaminated evaluation benchmarks with identified training-data overlap (continued).

Evaluation Benchmark	Potentially Overlapping Training Datasets
SSv2	Grounded-VideoLLM, InternVideo, LLaVA-Video-178k, Tarsier2-Recap-585K, VideoChat2-IT, VM22M
STAR	Grounded-VideoLLM, MVBench, M4-Instruct, SpecTemp-80K, TimePro, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k
StoryStream	COF-DATA, VideoEspresso
SumMe	TimeIT, TimePro, VTG-IT-120K
TACoS	E.T.Instruct 164K
TemporalBench	ActivityNet-Captions, Charades, COIN, MPII-MD, Oops, ActivityNet, ActivityNet-RTL, Charades-Ego, Charades-STA, DeVE-QA, E.T.Instruct 164K, Grounded-VideoLLM, LLaVA-OV, LLaVA-Video-178k, Tarsier2-Recap-585K, TempGCap, TimeIT, TimePro, VideoXum, VTG-IT-120K, VTimeLLM
TGIF-QA	MMAT
Tomato	CLEVRER, PerceptionTest, TGIF-QA, COF-DATA, Grounded-VideoLLM, MMAT, MVBench, SpecTemp-80K, VideoChat2-IT, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k
TUNA	MiraData, PerceptionTest, YouCook2, COF-DATA, Grounded-VideoLLM, LLaVA-OV, LLaVA-Video-178k, LongVid, LongVR, MVBench, SpecTemp-80K, TextVR, TimeIT, TimePro, VideoChat2-IT, VideoExplorer, VideoMarathon, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k, VideoEspresso
TVBench	Charades-STA, CLEVRER, FunQA, MoVQA, NTU RGB+D, PerceptionTest, STAR, Charades, COF-DATA, Grounded-VideoLLM, MVBench, M4-Instruct, SpecTemp-80K, TimeIT, TimePro, VideoChat2-IT, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k
TVGBench	ActivityNet-Captions, Charades-STA, Ego4D, HiREST, TaCoS, ActivityNet, ActivityNet-RTL, Charades, DeVE-QA, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, E.T.Instruct 164K, Grounded-VideoLLM, Haystack-Ego4D, LLaVA-OV, LLaVA-Video-178k, LongVid, MVBench, SpecTemp-80K, Tarsier2-Recap-585K, TempGCap, TimeIT, TimePro, VideoExplorer, VideoMarathon, VideoXum, VTG-IT-120K, VTimeLLM
TVQA	MVBench
TVSum	TimeIT, TimePro, VTG-IT-120K
VATEX	Kinetics, Causal-VidQA, COF-DATA, FineAction, Grounded-VideoLLM, InternVideo, LLaVA-Video-178k, ShareGemini, Tarsier2-Recap-585K, VideoChat2-IT, VM22M, VideoEspresso
VideoChatGPT	ActivityNet, ActivityNet-Captions, ActivityNet-QA, DPO, E.T.Instruct 164K, GCG, Grounded-VideoLLM, LLaMA-VID, LLaVA-OV, LLaVA-Video-178k, LongVR, ShareGPTVideo, SpecTemp-80K, Tarsier2-Recap-585K, TextVR, TimePro, VideoChatGPT-Plus, VideoChat2-IT, VideoExplorer, VideoMarathon
VCGBench-Diverse	HDVILA100M, YouCook2, COF-DATA, Grounded-VideoLLM, LLaVA-OV, LLaVA-Video-178k, LongVR, SpecTemp-80K, TextVR, TimeIT, TimePro, VideoChat2-IT, VideoExplorer, VideoMarathon, VideoEspresso
VCR-Bench	ActivityNet Captions, COIN, MLVU, NExT-QA, PerceptionTest, TVbench, VCGBench-Diverse, VideoEspresso, ActivityNet, ActivityNet-RTL, COF-DATA, DeVE-QA, E.T.Instruct 164K, Grounded-VideoLLM, Intent-QA, MVBench, M4-Instruct, SpecTemp-80K, TempGCap, TimeIT, TimePro, VideoChat2-IT, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k, VideoXum, VTG-IT-120K, VTimeLLM, CUVA, MovieChat-1K, SQA3D, StoryStream, VATEX, XD-Violence, Youcook2
VDC/VideoDC	Ego4D, Panda-70M, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Grounded-VideoLLM, Haystack-Ego4D, Koala-36M, LLaVA-OV, LLaVA-Video-178k, LongVid, SpecTemp-80K, Tarsier2-Recap-585K, VideoExplorer, VideoMarathon
VidCapBench	ActivityNet, NExT-QA, ActivityNet-Captions, ActivityNet-QA, DPO, E.T.Instruct 164K, GCG, Grounded-VideoLLM, Intent-QA, LLaVA-OV, LLaVA-Video-178k, LongVR, M4-Instruct, ShareGPTVideo, SpecTemp-80K, Tarsier2-Recap-585K, TextVR, TimePro, VideoChatGPT, VideoChat2-IT, VideoExplorer, VideoMarathon, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k
VideoEspresso	CUVA, MovieChat-1K, SQA3D, StoryStream, VATEX, XD-Violence, Youcook2, COF-DATA, Grounded-VideoLLM, LLaVA-OV, LLaVA-Video-178k, LongVR, SpecTemp-80K, Tarsier2-Recap-585K, TextVR, TimeIT, TimePro, VideoChat2-IT, VideoExplorer, VideoMarathon
VideoEval-Pro	MLVU
VideoHallucer	Activitynet-QA, Egoschema, MSRVTT, MSRVTT-QA, MSVD-QA, Next-QA, TGIF-QA, ActivityNet, Grounded-VideoLLM, Intent-QA, MMAT, M4-Instruct, VideoChat2-IT, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k
VideoMarathon	ActivityNet, Ego4D, MovieChat-1K, Panda-70M, YouCook2, ActivityNet-Captions, ActivityNet-QA, COF-DATA, DPO, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, E.T.Instruct 164K, GCG, Grounded-VideoLLM, Haystack-Ego4D, Koala-36M, LLaVA-OV, LLaVA-Video-178k, LongVid, LongVR, ShareGPTVideo, SpecTemp-80K, Tarsier2-Recap-585K, TextVR, TimeIT, TimePro, VideoChatGPT, VideoChat2-IT, VideoExplorer, VideoEspresso
VideoRewardBench	MVBench, VCGBench-Diverse, VSI-Bench, ActionBench, Charades-STA, CLEVRER, FunQA, Moments in Time V1, MoVQA, NTU RGB+D, PerceptionTest, STAR, TVQA, VLN-CE
VideoVista	Panda-70M, Grounded-VideoLLM, Koala-36M, VideoExplorer, VideoMarathon
VidHal	MVBench, PerceptionTest, ActionBench, Charades-STA, CLEVRER, FunQA, Moments in Time V1, MoVQA, NTU RGB+D, STAR, TVQA, VLN-CE, SpecTemp-80K, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k
VidOR	YFCC100M, LLaVA-Video-178k
VidSTG	VidOR, GCG, LLaVA-Video-178k
VidText	MLVU
VIM	EgoSchema, FunQA, NExT-QA, STAR, Grounded-VideoLLM, Intent-QA, MVBench, M4-Instruct, SpecTemp-80K, TimePro, VideoChat2-IT, VideoP2R-CoT-162K, Video-R1-CoT-165k, Video-R1-260k
VinoGround	VATEX, COF-DATA, Tarsier2-Recap-585K, VideoEspresso
VISTA-400K	FineVideo, InternVid, MiraData, OpenVid-1M, Panda-70M, Grounded-VideoLLM, InternVid-G, InternVid-VTime, InternVideo2, Koala-36M, LLaVA-Video-178k, LongVid, LongVR, ShareGemini, TimePro, TimeRFT, VideoExplorer, VideoMarathon, VTimeLLM
VITATECS	MSRVTT, VATEX, COF-DATA, Tarsier2-Recap-585K, VideoEspresso
VLN-CE	Matterport3D, MVBench
VMB	Ego4D, EPIC-KITCHENS, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Haystack-Ego4D, LLaVA-OV, LLaVA-Video-178k, LongVid, SpecTemp-80K, Tarsier2-Recap-585K, VideoExplorer, VideoMarathon
VSI-Bench	ScanNet
V-STaR	GOT-10K, TVQA+, VidSTG, GCG, MVBench, VideoChat-R1-18k
WorldQA	Ego4D, Epic-Kitchens, PVSG, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Haystack-Ego4D, LLaVA-OV, LLaVA-Video-178k, LongVid, SpecTemp-80K, Tarsier2-Recap-585K, VideoExplorer, VideoMarathon
WorldSense	FineVideo
XD-Violence	COF-DATA, VideoEspresso
X-LeBench	Ego4D, EgoIT-99K, Ego4D-HCap, Ego4D-livechat, Ego-QA, Haystack-Ego4D, LLaVA-OV, LLaVA-Video-178k, LongVid, SpecTemp-80K, Tarsier2-Recap-585K, VideoExplorer, VideoMarathon
YouCook2	COF-DATA, Grounded-VideoLLM, LLaVA-OV, LLaVA-Video-178k, LongVR, SpecTemp-80K, TextVR, TimeIT, TimePro, VideoChat2-IT, VideoExplorer, VideoMarathon, VideoEspresso

Table 7: Potentially contaminated evaluation benchmarks with identified training-data overlap (continued).