Vicinagearth

REVIEW

Open Access

Towards training-free long video understanding: methods, benchmarks, and open challenges

Jingren Liu[1,2†], Yun Wang[1,3†], Long Zhang[1,4†], Yiheng Wang[1], Shuning Xu[1,5], Ling Wang[1], Jiaqi Yan[1,6], Dell Zhang[1] and Xiangyu Chen[1*]

Abstract

The explosive proliferation of long-form video has exposed fundamental limitations in current AI systems, particularly in sustaining coherent temporal reasoning under stringent computational budgets. This survey synthesizes recent progress in long video understanding (LVU) with a deliberate focus on training-free methods, approaches that forgo retraining pretrained base models and instead construct dedicated LVU pipelines to maximally leverage their latent capabilities. We identify three core challenges that structure this landscape: (i) pervasive visual-token redundancy that inflates computation while contributing marginal information gain, (ii) constrained context windows that fragment temporal and semantic coherence, and (iii) the requirement for robust multi-modal reasoning across expansive temporal horizons. To address these challenges, existing solutions can be organized into three methodological paradigms. First, selection-based approaches target redundancy via semantically informed sampling and compression. Second, memory-enhanced architectures expand the effective temporal receptive field through hierarchical representations and streaming-aware mechanisms. Third, agent-based reasoning frameworks reconceptualize LVU as an active, iterative process driven by explicit goals and context-sensitive inference, enabling dynamic control over when and how video segments are processed. In parallel, we examine a growing specialization of evaluation protocols intended to assess LVU systems holistically. These protocols coalesce into three principal categories: general-purpose benchmarks that probe foundational, cross-task capabilities; reasoning-centric benchmarks that stress-test higher-order competencies, such as spatio-temporal abstraction, causal and physical reasoning, and compositional understanding; and egocentric benchmarks that emphasize first-person perception and action reasoning within dynamically evolving real-world environments. Collectively, these benchmarks illuminate both the representational fidelity and the reasoning depth required for robust deployment. Our findings indicate a pivotal inflection in the field: ad hoc, engineering-driven heuristics are giving way to principled, end-to-end frameworks that integrate scalable computation with cognitively grounded reasoning. Rather than merely amplifying base-model capacity, this shift advances a coherent agenda centered on pragmatic, auditable, and resource-aware LVU pipelines atop pretrained models, thereby extending LVU to professional analytics, embodied human–AI interaction, and safety-critical decision making.

Keywords Training-free long video understanding, Multimodal large language models, Benchmarks

† Jingren Liu, Yun Wang and Long Zhang contributed equally to this work.

*Correspondence:
Xiangyu Chen
chxy95@gmail.com
Full list of author information is available at the end of the article

Springer

© The Author(s) 2025. **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

1 Introduction

The exponential surge of long-form video across digital platforms has fundamentally reshaped the problem space of visual understanding [1–3]. From educational lectures and live-streamed events to surveillance repositories and multimedia analytics, contemporary systems are expected to operate over temporally extended sequences that span hours, or even days, while preserving semantic coherence and computational feasibility [4]. With more than 500 hours of video uploaded to YouTube every minute, the sheer scale and continuity of unstructured footage demand frameworks capable of reasoning over long temporal horizons without incurring prohibitive costs.

Historically, long video understanding (LVU) [5] has been dominated by supervised pipelines that require large labeled datasets, task-specific architectural tuning, and repeated retraining cycles [6–8]. While effective in narrowly scoped settings, these approaches exhibit poor scalability, limited domain generalization, and significant deployment overhead. Their dependence on task-specific fine-tuning makes them brittle under distribution shifts, prone to catastrophic forgetting, and ill-suited for open-domain reasoning [9–12]. In contrast, an emerging line of work advocates training-free methodologies that operate atop frozen, pretrained backbones without parameter updates. Crucially, the contribution of this paradigm is not merely ease of deployment, but the formulation of a pragmatic and verifiable LVU pipeline: one that systematically orchestrates inference-time modules, retrieval, compression, scheduling, and reasoning controllers, around fixed models to directly confront the intrinsic challenges of long video analysis. By decoupling representational capacity from training budgets, such pipelines enable resource-aware adaptation, facilitate transparent diagnostics and error attribution, and support rapid, task-driven reconfiguration with strong auditability and reproducibility. The momentum toward training-free architectures is propelled by three structural obstacles that end-to-end schemes continue to struggle with: (i) suppressing massive visual-token redundancy while preserving representational fidelity, (ii) maintaining temporal and contextual coherence under constrained context windows, and (iii) executing robust, multi-modal reasoning across expansive temporal spans. Rather than seeking to further amplify base-model capacity in isolation, the proposed pipeline-centric perspective reframes LVU as an engineering of interfaces and control flows that make long-horizon inference tractable, dependable, and genuinely aligned with real-world LVU requirements.

First, the problem of visual-token redundancy becomes pronounced in long video settings: frame-level feature extraction across thousands of frames results in computational overload that quickly exceeds hardware constraints [13–16]. Second, most existing architectures impose fixed context windows that fragment video continuity, making it difficult to capture long-range dependencies and hierarchical temporal structure [7, 17]. Third, long video comprehension entails temporal reasoning tasks, such as causal reasoning, event prediction, and narrative disambiguation, that surpass low-level visual perception and require reasoning mechanisms capable of abstraction and multi-step reasoning [18–20].

Training-free solutions (as illustrated in Fig. 1) address these challenges through three complementary paradigms, each leveraging frozen foundation models at inference time. Selection-based methods reduce input complexity by performing content-aware sampling, generative frame ranking, or logic-driven token pruning. These techniques identify semantically salient regions without retraining, enabling scalable input compression [21, 22]. Memory-enhanced frameworks introduce dynamic memory structures, such as hierarchical video segmentation [23], temporal caches [24], and continuous-time memory systems [25], to extend contextual reasoning across long durations without increasing inference cost linearly. Agent-based reasoning approaches represent a more recent advancement: instead of processing videos passively, these systems orchestrate multiple pre-trained components, vision encoders, language models, retrievers, and planners, via iterative protocols that emulate high-level cognitive strategies such as exploration, reflection, and plan adjustment [26, 27].

This reorientation from optimization-intensive pipelines to inference-driven architectures signals a paradigmatic shift in LVU. Rather than updating parameters to accommodate each new task or domain, training-free systems rely on structured orchestration of capabilities embedded in foundation models. By combining efficient token selection, temporal memory mechanisms, and multi-agent collaboration, these approaches unlock robust temporal reasoning while preserving cross-domain generalization and real-time deployability. As such, training-free LVU emerges as a scalable, cognitively aligned, and computationally sustainable paradigm, offering a compelling alternative to classic model-centric training regimes in the era of ever-growing visual data.

While training-free methodologies offer notable advantages in scalability, generalization, and deployment efficiency, their evaluation poses distinct challenges. Unlike traditional approaches that benefit from task-specific fine-tuning and tailored dataset alignment, training-free systems rely exclusively on pre-trained representations and inference-time reasoning. This necessitates more rigorous evaluation protocols: benchmarks must not only measure foundational understanding [28–30] but also probe models' reasoning capabilities under constraints

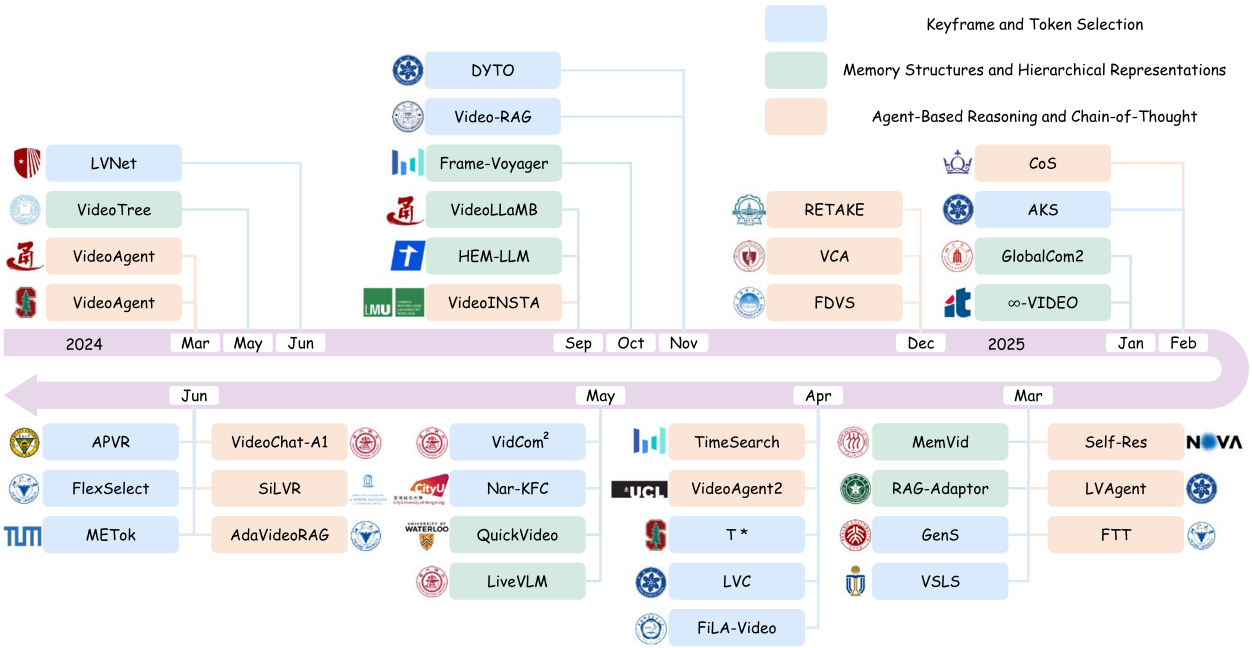


Fig. 1 An overview of the timeline of existing training-free methods for long video understanding

such as frozen weights, limited context, and dynamic reasoning scenarios.

To meet these demands, the research community has developed a suite of benchmark ecosystems (as illustrated in Fig. 2) tailored to different aspects of LVU. These benchmarks serve as critical instruments for diagnosing system limitations, guiding method selection, and aligning future development with real-world deployment scenarios. We categorize them into three complementary domains.

General LVU benchmarks serve as the foundation for evaluating training-free models' capacity to perform basic comprehension across diverse domains, modalities, and video lengths. Systems operating without supervised adaptation must demonstrate stable performance across wide distributions and temporal scales. MVBench [31]

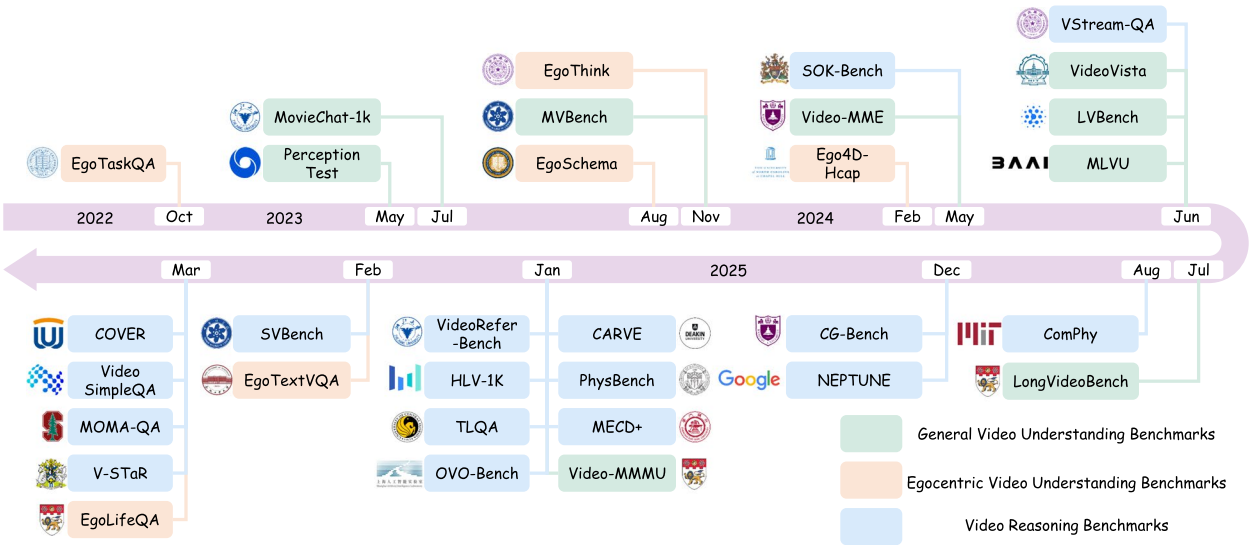


Fig. 2 An overview of the development timeline of existing benchmarks for long video understanding

introduces a temporally structured task taxonomy, from visual recognition to abstract reasoning, exposing sharp performance drop-offs in Video-LLMs as tasks increase in complexity. Similarly, Perception Test [32] provides a diagnostic view across six reasoning dimensions using controlled real-world videos. VideoVista [33] and Video-MME [34] extend this by scaling to tens of thousands of questions across multimodal categories, enabling stress-testing of frozen models' generalization limits. Of particular relevance to training-free approaches are benchmarks that target long-horizon scalability without retraining. MovieChat [35] and LVBench [36] evaluate performance across hour-long and even multi-hour videos, revealing steep degradation in contextual coherence. MLVU [7] and LongVideoBench [6] further challenge memory-constrained models with interleaved video-text formats and multi-task demands. These findings validate the need for efficient selection and compression techniques that minimize token load while preserving semantic fidelity, core goals of training-free LVU pipelines. Beyond perception, training-free systems must also support high-level reasoning under limited context and static parameters. This necessitates benchmarks that evaluate temporal abstraction, causal reasoning, and symbolic integration in zero-shot conditions. Spatio-temporal reasoning tasks, as in VideoRefer [37], V-STaR [38], and TimeLogic [39], reveal weaknesses in capturing relational dependencies without explicit tuning. Streaming tasks such as SVBench [40] assess whether models can maintain temporal coherence across evolving contexts, essential for inference-time memory routing. Causal and physical reasoning benchmarks (e.g., CARVE [18], MECD+ [19], PhysBench [41]) probe deeper into dynamic understanding. Training-free agent-based models must infer causes, latent properties, and counterfactual outcomes without feedback-based learning. These benchmarks are particularly informative for evaluating iterative reasoning strategies that emulate plan-adjust or reflect-explore behaviors. Finally, knowledge-grounded datasets like Video SimpleQA [42], CG-Bench [43], and SOK-Bench [44] assess models' ability to align retrieved knowledge with temporally grounded visual inputs. For training-free architectures, which cannot acquire new knowledge, these evaluations reveal the limits of retrieval-augmented and context-injected LVU pipelines. Egocentric scenarios introduce embodied challenges not present in static third-person footage, demanding fine-grained spatial grounding, dynamic viewpoint tracking, and interactive reasoning. Benchmarks like EgoSchema [45] and EgoTaskQA [46] evaluate long-range temporal modeling and action-cause prediction under personalized, noisy conditions, areas where memory-augmented training-free systems often falter due to a lack of semantic persistence. Scene text and goal-oriented reasoning benchmarks such as Ego-TextVQA [47] and EgoThink [48] expose critical gaps in cross-modal alignment, especially when models must synthesize visual context, OCR text, and implicit user goals without gradient updates. EmbodiedEval [49] pushes this further by embedding evaluation within simulation environments, requiring multi-step planning and object interaction under realistic task prompts. These benchmarks are key to assessing whether training-free agents can generalize sensorimotor inference without in-situ tuning.

Together, these benchmark categories form a tiered evaluation scaffold for training-free LVU. General benchmarks highlight the need for efficient selection and compression; reasoning benchmarks emphasize the value of agent-style orchestration for abstract cognition; and egocentric benchmarks stress-test embodied, context-rich interaction. By systematically exposing failure modes across perception, reasoning, and embodiment, these evaluations provide a rigorous basis for guiding the next generation of training-free video architectures.

2 Training-free LVU methods

2.1 Memory Structures and Hierarchical Representations

With the explosive growth of ultra-long video data, accurately extracting critical content while maintaining computational efficiency has emerged as a central challenge in the development of large-scale video models. Beyond a variety of keyframe and key token selection strategies, approaches based on memory structures and hierarchical representations have also proliferated. These methods can be broadly categorized into three directions: (1) adaptive hierarchical architectures that dynamically organize video information; (2) memory-augmented and token-compression schemes that enhance model capacity and speed; and (3) online streaming frameworks that leverage retrieval mechanisms for real-time interaction. Representative frameworks from each of these directions are reviewed below, followed by a discussion of promising future trends.

2.1.1 Adaptive hierarchical methods

Adaptive hierarchical methods [23, 24, 50] employ multi-level designs to dynamically filter and organize video content, thereby balancing information completeness against redundancy suppression. For example, VideoTree [24] adopts a three-stage pipeline: adaptive breadth expansion, relevance-guided depth expansion, and hierarchical tree representation, to extract and query the most relevant key frames in real time, building a multilevel tree that effectively mitigates redundant content and structural omissions. HEM-LLM [23] introduces an event-based adaptive sequence segmentation, partitioning

a long video into multiple units of logical events and establishing multigranularity memories both within and between events to reinforce long-term dependencies and the association of information between events.

2.1.2 Memory-augmented approaches

Memory-augmented and token-compression approaches [25, 51, 52] focus on boosting long-video reasoning capacity and speed under limited computational budgets. GlobalCom2 [51] introduces the concept of "global thumbnail guidance for local crop token compression": a global thumbnail is first used to assess the importance of `[CLS]` tokens, and then a compression budget is adaptively allocated to local regions based on their information richness, yielding substantial reductions in visual tokens while preserving high performance. Inspired by human "sticky memory", ∞-VIDEO [25] employs a continuous-time long-term memory integration mechanism that dynamically allocates attention granularity within a single video pass, simulating memory evolution through continuous attention and significantly improving performance on multiple video-QA benchmarks without additional training.

2.1.3 Real-time streaming applications

For real-time streaming applications, methods in this category [53–55] achieve low latency and high throughput through joint algorithmic and architectural optimization. LiveVLM [53] builds a "streaming KV-cache + dynamic retrieval" framework that generates and compresses key/value tensors in real time, retaining long-term video details while eliminating redundancy. QuickVideo [54] employs QuickCodec for parallel CPU decoding of key frames, QuickPrefill for intra-group pre-filling and KV-cache pruning, and overlapping CPU/GPU execution to drastically reduce end-to-end processing delay. RAG-Adapter [55], evaluated on long-video benchmarks, replaces uniform sampling with question-driven frame retrieval and introduces metrics such as group contrastive learning (GCL), average similarity score (ASS), and necessary information frames (NIF) to enable precise evaluation and reliable performance gains.

In general, LVU is converging along the axes of "adaptive hierarchy, +, memory enhancement, +, efficient compression/streaming." Adaptive hierarchical structures strike a balance between information extraction and contextual coherence; memory and compression techniques reduce the dependency on large-scale training and resources; and streaming with system level co-optimization meets the efficiency demands of real-time applications. Future research may further explore cross-event multimodal associations, interpretable memory retrieval, and edge-cloud collaborative inference to support intelligent video understanding in ever-longer and more diverse scenarios.

2.2 Keyframe and token selection

In recent years, three core challenges have emerged in LVU: visual-token redundancy, limited memory and context length, and the efficient extraction of critical information. To address keyframe and key-token selection, the community has developed three main strategies: (1) seamlessly integrated compression methods that exploit content awareness or multi-stage scheduling to reduce visual tokens; (2) retrieval-based methods, employing dual-granularity search or spatio-temporal reconstruction, to locate and preserve essential information; and (3) keyframe selection and ranking schemes driven by diverse scoring or logical modeling techniques, which extract and query the most relevant visual inputs at the frame level.

2.2.1 Compression methods

Seamlessly integrated compression methods [56] perform fine-grained token reduction via content-aware or multi-stage mechanisms. For example, VidCom^2 [57] adopts a two-stage adaptive compression: it first computes a distinctiveness score for each frame based on inter-frame similarity and allocates a token budget via softmax; it then ranks all visual tokens by combining intra-frame and cross-video distinctiveness and applies Top-K pruning under the given budget. METok [14] proposes an event-aware, multi-stage token compression pipeline: during encoding, it segments continuous events by cosine similarity and identifies key frames; during prefill, it hierarchically prunes tokens by attention strength and event importance; and during decoding, it discards low-impact tokens in the KV cache to minimize floating-point operations (FLOPs) and memory usage. DYTO [13] realizes efficient, zero-shot compression by hierarchically clustering at the frame level into dynamic key-token groups, then adaptively merging tokens from coarse to fine through a binary merging process.

2.2.2 Retrieval-based key-information localization approaches

Retrieval-Based key-information localization approaches [58, 59] leverage dual-granularity or spatio-temporal reconstruction strategies to efficiently select and retain the most relevant frames and tokens. APVR [58] first performs pivot frame retrieval (PFR) and scoring, then applies attention-based token filtering within the selected frames, and finally fuses spatio-temporal semantic confidence with query-aware attention scores. T* [59] reformulates temporal retrieval in long videos as a spatial search problem, adaptively adjusting search granularity

across spatio-temporal dimensions to significantly enhance keyframe localization under extreme frame-budget constraints.

2.2.3 Multi-scoring and logic-driven frame selection methods

Multi-Scoring and Logic-Driven Frame Selection methods [21, 22, 60, 61] employ multiple scoring functions or logical reasoning to precisely extract and query the most pertinent visual inputs at the frame level. AKS [21] proposes an adaptive keyframe sampling algorithm that balances relevance and coverage by computing frame-prompt matching scores and recursively applying a "judge-and-split" strategy. GenS [60] uses a long-context VideoLLM to generate frame relevance predictions conditioned on user instructions, guiding sampling. VSLS [62] constructs four types of logical triples, spatial co-occurrence, temporal proximity, attribute dependency, and causal order, and iteratively optimizes frame distribution based on logical confidence for context-aware keyframe identification. Nar-KFC [63] formulates keyframe selection as an integer quadratic programming problem to jointly optimize relevance and diversity, solves it approximately via a customized greedy strategy, and inserts coherent textual narratives generated from non-key frames to mitigate temporal discontinuities. LVNet [64] employs a hierarchical selector, temporal scene clustering (TSC), coarse keyframe detector (CKD), and fine keyframe detector (FKD), and leverages a small amount of subtitle supervision to support long-video question answering. Overall, these methods have evolved from "training-free, plug-and-play" compression schemes to "content-aware, multi-stage scheduling" for dynamic token management, and further to "multi-granularity retrieval and logic-driven" keyframe extraction. They continually advance toward greater generality, interpretability, and reasoning efficiency, laying a solid foundation for the application of multimodal large models in LVU.

2.3 Agent-based reasoning and chain-of-thought

Recent advancements in LVU have leveraged the synergistic integration of Agent-Based Reasoning and Chain-of-Thought (CoT) prompting. These paradigms address the inherent challenges of extended temporal sequences, such as capturing fine-grained spatiotemporal dynamics, mitigating semantic redundancy, and modeling long-range dependencies, by transforming passive multimodal processing into active, iterative reasoning processes.

2.3.1 Agent-based approaches

These methods frame video understanding as an agentic process, where LLMs/MLLMs collaborate, explore, or iterate to build comprehension. Agents often use cycles of selection, action, and reflection to handle video length efficiently. Remarkably, LVAgent [26] proposes a multi-round framework with select-perceive-act-reflect steps for MLLM agents to dynamically collaborate on long videos, improving understanding without fixed sampling. Videoagent [65] uses an LLM as a central agent to iteratively identify and compile key information from long videos, answering queries via progressive refinement. VCAgent [66] introduces a curiosity-driven agent with self-exploration capabilities, autonomously navigating video segments to build a comprehensive understanding of complex sequences. VideoAgent [67] combines LLMs and vision-language models with a unified memory mechanism to capture long-term temporal relations, addressing challenges in extended videos. VideoAgent2 [27] builds on prior agents with an uncertainty-aware CoT process, using plan-adjust modes to progressively acquire information and enhance reasoning in long videos.

2.3.2 CoT and reasoning paradigms

These emphasize structured reasoning, such as shot-based chains or hierarchical searches, to mimic human-like step-by-step analysis and reduce hallucinations in long contexts. VideoINSTA [68] is a zero-shot framework using event-driven temporal reasoning and content-driven spatial reasoning, enabling LLMs to infer spatiotemporal information without training. VideoChat-A1 [69] designs a "chain-of-shot" paradigm with multi-round shot division and MLLM dialogues to progressively understand user queries in long videos, achieving SOTA on Video-MME (77.0%) and EgoSchema (70.1%). CoS [70] introduces shot-sequence structured prompting to guide LLMs in understanding long videos through sequential CoT. SiLVR [71] is a modular, training-free framework that transforms videos into language representations (using captions, subtitles) then applies LLM reasoning with adaptive token reduction, excelling on Video-MME (long) and Video-MMMU. Self-ReS [72] incorporates self-reflection for LVLMs to self-assess and adjust during long video processing, using sparse attention for dynamic sampling, yielding 46% faster inference. FTT [73] uses bin-wise sampling for diverse predictions and self-reward (frequency, confidence, reasoning scores) to select finals, improving MLLM performance across 7 datasets. Video-EM [74] proposes episodic memory representations to present the key events and use the Chain-of-Thought strategy to filter unreliable events.

2.3.3 Retrieval-augmented methods

RAG-based methods augment models with external retrieval to fetch relevant video contexts, often adaptively, for cost-effective long video comprehension.

Video-RAG [75] is a training-free pipeline with visually aligned auxiliary text to enhance cross-modal alignment and provide extra info beyond visuals. AdaVideoRAG [76] can dynamically adapt retrieval granularity via query complexity using an intent classifier and omni-knowledge indexing (text, visuals, graphs), introducing the HiVU benchmark for efficiency gains.

Recent progress in LVU has led to a growing suite of training-free solutions, encompassing keyframe selection, hierarchical memory structures, token compression, and agent-based reasoning. These methods address the core challenges of temporal scale and computational overhead through complementary strategies in compression, retrieval, and inference-time coordination. While memory-augmented and hierarchical designs focus on scalability, agent-based and CoT-style reasoning promote active, structured understanding over passive processing. Retrieval-augmented pipelines further extend adaptability by interfacing frozen models with external, context-aware memory. Together, these advances reflect a clear shift toward modular, interpretable, and inference-efficient architectures for open-domain, long-horizon video analysis. Yet, their practical trade-offs remain underexplored. To this end, the next subsection provides a systematic comparison of representative methods across standardized benchmarks, highlighting performance trends and revealing the operational boundaries of current training-free LVU paradigms.

2.4 Comparative evaluation of training-free LVU methods

We systematically evaluate the performance of various training-free video understanding models [58, 59] and their baseline methods [77–79] on the Video-MME dataset [34] under different video lengths (short, medium, and long) and input configurations, as shown in Table 1. The evaluation focuses on analyzing the effects of large language model size (LLM-Size), number of input frames (Frames), and additional optimization strategies on model performance, while also revealing performance trends across models under varying task difficulty levels.

2.4.1 Overall performance comparison

In terms of the *Overall* metric, **Qwen2.5VL-72B+FlexSelect** [80] achieves the highest performance with a score of 74.4, outperforming Qwen2.5VL-72B (73.4) and LLaVA-Video-72B+FlexSelect (73.1). These results indicate that integrating large-scale multimodal language models with the *FlexSelect* strategy can effectively enhance the model's ability to comprehend complex video content and improve overall task performance.

2.4.2 Impact of LLM size

Experimental results demonstrate that increasing the parameter scale of large language models (LLMs) generally leads to improved video understanding performance. Taking *VILA** + *Frame-Voyager* [61] as an example, when the model size increases from 8B to 34B, the overall score improves from 50.5 to 60.0 (+9.5). Specifically, the short-task score rises from 60.3 to 70.3 (+10.0), the medium-task score from 47.3 to 58.3 (+11.0), and the long-task score from 43.9 to 51.2 (+7.3). A similar trend is observed in the *LLaVA-OneVision* [79] series, where the 72B model achieves a significantly higher overall score of 66.3 compared to 56.5 for the 7B variant, further validating the advantage of larger models in capturing long-term temporal dependencies and complex semantic structures in video content.

2.4.3 Impact of input frames

The impact of input frame count (Frames) on model performance varies by model. For the GPT4o series [81], the overall performance with 32 frames (62.5) is better than that with 8 frames (61.7), especially with a noticeable improvement in short tasks (71.4 vs 69.6). For LLaVA-OneVision-72B, the overall performance with 32 frames (66.3) is significantly higher than that with 8 frames (55.9), suggesting that sufficient input frames provide richer temporal information for large models. Additionally, the METok model [14] achieves much higher scores in medium/long tasks with 128 frames (46.6/52.4) compared to 32 frames (36.4/42.7), validating the importance of long video inputs for long-term temporal tasks.

2.4.4 Effect of additional optimization strategies

Various additional strategies (e.g., T* [59], VSLS [62], FlexSelect [80], Frame-Voyager [61]) consistently improve the performance of base models. For instance:

- LLaVA-OneVision-72B (32 frames) with T* achieves an overall score of 68.3, up from 66.3 (+2.0), with the medium-task score increasing from 62.2 to 66.6 (+4.4);
- Qwen2.5 VL-7B with FlexSelect improves the medium-task score from 55.6 to 59.3 (+3.7) and the overall score from 65.4 to 68.2 (+2.8);
- VILA*−34B combined with Frame-Voyager increases the overall performance from 58.3 to 60.0 (+1.7), indicating that temporal frame selection strategies effectively reduce redundant information and focus on key content.

Table 1 Comparison of different models and configurations on Video-MME

Model	LLM-size	Frames	Short	Medium	Long	Overall
GPT4o	-	8	69.6	60.2	55.2	61.7
GPT4o + (T*)	-	8	68.9	61.2	55.2	61.8
GPT4o + VSLS	-	8	68.2	60.7	56.9	61.9
GPT4o	-	32	71.4	61.0	55.2	62.5
GPT4o + (T*)	-	32	72.6	61.6	55.2	63.1
GPT4o + VSLS	-	32	71.9	61.9	55.2	63.0
VIDEOTREE	-	-	-	-	-	54.2
INTERNVL2.5	78B	8	55.9	55.5	52.6	54.6
INTERNVL2.5 + VSLS	78B	8	59.0	57.5	57.7	58.1
InternVL2	7B	16	67.7	49.9	44.4	54.0
InternVL2	20B	16	65.8	50.3	45.4	53.9
InternVL2	34B	16	72.0	59.1	52.6	61.2
InternVL2	8B	-	62.1	48.2	45.2	51.9
InternVL2 + KFC	8B	-	64.3	49.6	46.1	53.1
InternVL2 + Nar-KFC	8B	-	67.2	54.7	47.1	56.3
InternVL2.5	8B	-	-	-	52.8	64.2
InternVL2.5 + FlexSelect	8B	-	-	-	58.1	67.0
InternVL2.5 + FlexSelect-Lite	8B	-	-	-	57.9	67.2
LLaVA-One-Vision	7B	-	64.0	52.1	43.8	53.3
LLaVA-One-Vision + Frame-Voyager	7B	-	67.3	56.3	48.9	57.5
LLaVA-One-Vision	7B	32	-	-	-	56.5
LLaVA-One-Vision w/AKS	7B	32	-	-	-	58.4
LLaVA-OneVision	72B	8	59.6	55.5	52.6	55.9
LLaVA-OneVision + (T*)	72B	8	61.7	57.5	57.7	59.0
LLaVA-OneVision	72B	32	76.7	62.2	60.0	66.3
LLaVA-OneVision + (T*)	72B	32	77.5	66.6	61.0	68.3
LLaVA-Video	7B	64	-	-	-	64.4
LLaVA-Video w/AKS	7B	64	-	-	-	65.3
LLaVA-Video + ReTaKe	7B	-	-	-	55.4	64.0
LLaVA-Video + FlexSelect	7B	-	-	-	59.8	68.9
LLaVA-Video + FlexSelect-Lite	7B	-	-	-	58.3	68.3
LLaVA-Video	72B	-	-	-	61.9	70.0
LLaVA-Video + FlexSelect	72B	-	-	-	66.1	73.1
LLaVA-NeXT-Video	34B	-	61.7	50.1	44.3	52.0
LLoVi	-	-	-	-	-	48.8
Qwen2-VL	7B	32	-	-	-	57.6
Qwen2-VL w/AKS	7B	32	-	-	-	59.9
Qwen2-VL + ReTaKe	7B	-	-	-	56.2	63.9
Qwen2.5-VL	7B	-	-	-	55.6	65.4
Qwen2.5-VL + FlexSelect	7B	-	-	-	59.3	68.2
Qwen2.5-VL + FlexSelect-Lite	7B	-	-	-	58.6	67.4
Qwen2.5 VL	72B	-	-	-	63.9	73.4
Qwen2.5 VL + FlexSelect	72B	-	-	-	66.9	74.4
VideoChat2	Mistral-7B	32	-	37.9	38.0	42.1
∞-VideoChat2 (no LTM)	Mistral-7B	128	-	39.6	38.8	42.3
∞-VideoChat2 (uniform)	Mistral-7B	128	-	40.0	38.8	42.4
∞-VideoChat2 (sticky)	Mistral-7B	128	-	40.2	38.9	42.4
VideoLLaMA2	8×7B	-	-	-	-	47.9
VILA*	8B	-	57.8	44.3	40.3	47.5

Table 1 (continued)

Model	LLM-size	Frames	Short	Medium	Long	Overall
VILA*+ Frame-Voyager	8B	-	60.3	47.3	43.9	50.5
VILA*	34B	-	67.9	56.4	50.4	58.3
VILA*+ Frame-Voyager	34B	-	70.3	58.3	51.2	60.0
VITA	8×7B	-	65.9	52.9	48.6	55.8
FastV	-	-	65.0	53.8	47.0	55.3
PDrop	-	-	64.7	53.1	48.7	55.5
SparseVLM	-	-	68.4	55.2	48.1	57.3
DyCoke	-	-	61.1	48.6	43.2	51.0
VidCom^2	-	-	69.8	56.4	49.4	58.6
METok	0.5B	32	-	-	36.4	42.7
METok	7B	128	-	-	46.6	52.4

boldface indicates the backbone used by the corresponding training-free method

2.4.5 Task difficulty and model performance

All models achieve significantly higher scores on **short tasks than medium and long tasks**, indicating that LVU remains a major challenge. For example, GPT4o (32 frames) scores 71.4 on short tasks but only 55.2 on long tasks (a 16.2-point gap). Even the high-performing *LLaVA-OneVision-72B + T* [59] model demonstrates its advantage primarily on short tasks—achieving a short-task score of 77.5, whereas its long-task score is only 61.0. This result suggests that improving models' ability to model long-term temporal dependencies should be a core direction for future research.

2.5 Lightweight multimodal fusion strategies

Recent advances in lightweight multimodal fusion have demonstrated that strong video understanding performance can be preserved while substantially reducing computational overhead. These methods follow three key design principles that collectively enable efficient deployment.

Parameter-efficient adaptation. Several works achieve competitive performance by training only a small fraction of model parameters. UniAdapter [82] introduces cross-modal adaptation modules for vision-language models, tuning merely 1–2% of parameters while matching full fine-tuning performance on downstream tasks. Similarly, FLoRA [83] extends low-rank adaptation to incorporate new modalities for speech detection, achieving significant error rate reduction with minimal trainable weights and robustness to modality dropout.

Dynamic computation allocation. Adaptive inference strategies further reduce computational demands by selectively activating model components. Polymorph [84] dynamically activates lightweight adapters per video frame based on label co-occurrence patterns, yielding 40% energy savings and improved mAP compared to static baselines. This approach demonstrates that temporal redundancy in video streams can be effectively exploited for efficiency gains.

Modular fusion architectures. Contemporary methods increasingly adopt modular designs that decouple representation learning from fusion. MMER-LMF [85] employs decision-level fusion of video and text emotion scores, achieving high accuracy with reduced training complexity. MoE-LoRA [86] adapts the Segment Anything Model by freezing core weights and training only modality-specific LoRA experts, maintaining low inference overhead. FuseMix [87] combines frozen unimodal encoders with lightweight fusion adapters and latent space augmentation, enabling multimodal reasoning on resource-constrained hardware.

While these parameter-efficient approaches show promise, several challenges remain unaddressed. Current methods often sacrifice fine-grained temporal alignment and cross-modal synchronization—critical factors for dynamic video understanding. Moreover, existing evaluations predominantly focus on accuracy metrics while overlooking deployment-critical factors such as latency, energy consumption, and memory footprint. Future work should establish comprehensive benchmarks that jointly evaluate model performance and practical deployment constraints, particularly for real-time video applications where both accuracy and efficiency are paramount.

2.6 Practical deployment value vs challenges

While training-free long-video understanding methods show strong potential, their real-world deployment is constrained by three practical factors: edge-device resources, latency requirements, and deployment costs.

Edge-device constraints. Resource-limited devices impose strict bounds on compute, memory, and power. Prior work [88] characterizes these limits and the

heterogeneity of edge hardware, which impede direct deployment of large models. Practical mitigation strategies follow a triad of optimizations: data-level (cleaning, compression, augmentation), model-level (pruning, quantization, distillation), and system-level (framework and hardware acceleration), which together enable feasible multimodal inference on edge platforms [89].

Latency-critical applications. Interactive scenarios (e.g., AR/VR) demand low end-to-end latency, exposing bottlenecks in sensing, transmission, and inference. Utility-aware load shedding—selectively discarding low-value frames—has been shown to reduce compute and network load while retaining essential temporal information, offering a practical trade-off between efficiency and accuracy [90].

Scalability and deployment costs. Beyond per-device compute, large-scale deployment incurs bandwidth, multimodal-fusion, and maintenance costs. Task-oriented communication that compresses and transmits only task-relevant features can substantially lower communication overhead and simplify deployment pipelines [91].

Despite these challenges, edge deployment delivers tangible benefits (reduced latency, improved privacy, offline capability), motivating continued research on joint data–model–system co-design [92]. Systematic optimization across the model lifecycle—from resource-aware design and pre-deployment preparation to runtime inference tuning—remains a practical path toward deployable long-video understanding systems [88].

3 LVU benchmarks

To enhance clarity, we organize benchmarks into three complementary categories: (i) general-purpose, (ii) reasoning-oriented, and (iii) egocentric. This structure is mirrored in the following subsections, each with its own summary and transition.

3.1 General LVU benchmarks

The rise of training-free LVU paradigms, spanning selection-based, memory-augmented, and agent-driven methods, calls for evaluation protocols that rigorously test core video understanding without task-specific tuning. In contrast to supervised approaches, these systems must generalize across domains and temporal complexities using only frozen representations and inference-time reasoning, prompting the design of benchmarks tailored to assess their foundational capabilities (see Table 2).

3.1.1 Core generalization evaluation

Training-free LVU systems fundamentally rely on the generalization capabilities embedded within frozen foundation models, necessitating comprehensive evaluation of zero-shot transfer performance to uncover system limitations and inform architectural design. MVBench [31] and Perception Test [32] provide essential diagnostic frameworks, systematically evaluating model behavior across temporal perception tasks and cognitive domains. MVBench's 20-task benchmark reveals consistent degradation in performance as task complexity increases: models achieve 85–88% accuracy on low-level visual perception tasks but drop to 30–40% on high-level reasoning, particularly for models lacking extensive instruction tuning. Perception Test, consisting of 11,609 real-world videos, highlights fundamental limitations in temporal memory and consistency. Frozen models exhibit significant degradation on tasks requiring long-range information integration, with notably reduced accuracy on longer video clips, suggesting current memory mechanisms

Table 2 Overview of general video understanding benchmarks

Benchmark	Year	Video length	Dataset scale	Focus area	Venue
Core Generalization Evaluation					
MVBench	2024	Avg. 16 sec	20 temporal tasks, 4 K QA pairs	Temporal reasoning	CVPR
Perception Test	2023	Avg. 23 sec	11.6K videos, 38 K multiple-choice QA pairs, 6 K grounded QA pairs	Multi-capability QA	NeurIPS
VideoVista	2024	Avg. 131 sec	3.4K videos, 14 video categories, 25 K QA pairs,	Narrative understanding	arXiv
Video-MME	2024	Avg. 1018 sec	6 main video domains, 30 video subfields, 0.9K videos, 2.7K QA pairs	Cross-modal generalization	CVPR
Scalability Assessment					
MovieChat-1K	2024	Avg. 564 sec	1 K video clips, 3 K global QA pairs, 10 K breakpoint QA pairs	Dialogue QA	CVPR
LVBench	2024	Avg. 4101 sec	103 videos, 6 QA tasks, 1549 QA pairs	Extreme long video	ICCV
MLVU	2024	Avg. 930 sec	1,730 videos, 3,102 QA pairs	Genre robustness	CVPR
LongVideoBench	2024	Avg. 673 sec	17 fine-grained categories, 3,763 videos, 6,678 QA pairs	Referring QA	NeurIPS
Specialized Domain Understanding					
Video-MMMU	2025	Avg. 506 sec	6 professional disciplines, 30 subjects, 300 videos, 900 QA pairs	Expert-level videos	arXiv

are inadequate for extended temporal dependencies. To complement these evaluations, VideoVista [33] and Video-MME [34] conduct cross-domain assessments, offering insights into the generalization boundaries of video foundation models. VideoVista's 25,000 curated questions across 14 diverse video categories reveal substantial variance in model performance across domains, underscoring the need for adaptive keyframe selection and domain-aware compression strategies. Video-MME, based on an expert-reviewed framework, analyzes extensive YouTube video data, showing that the consistency of training-free systems is largely confined to in-distribution domains. This observation exposes core constraints in generalization, especially in agent-based reasoning scenarios. Overall, these findings reveal that training-free LVU architectures face intrinsic trade-offs between generalization breadth and reasoning specialization, calling for selection strategies that preserve cognitive capacity while minimizing computational complexity.

3.1.2 Scalability assessment

Extended temporal scale evaluation provides critical insights into the operational boundaries where selection strategies, memory compression, and agent coordination mechanisms reach their effectiveness limits. MovieChat-1K [35] and LVBench [36] establish foundational stress-testing capabilities across hour-long to extreme temporal scales, revealing systematic performance degradation patterns that define the boundaries of training-free approaches. MovieChat's dense-to-sparse memory evaluation demonstrates that training-free compression frameworks achieve consistent performance on hour-long videos while enabling significant token reduction, informing critical design choices for hierarchical modeling. LVBench's evaluation across long-duration video tasks indicates performance degradation with increasing temporal scale, particularly beyond 2-hour contexts, although exact scaling trends remain subject to further validation. These patterns delineate practical constraints for long-range modeling and keyframe selection. MLVU [7], and LongVideoBench [6] address complementary scalability issues related to multi-task interference, interleaved processing, and hierarchical memory evaluation. MLVU highlights limitations of current architectures during task switching within continuous video contexts, revealing performance degradation when coordinating multiple cognitive processes without adaptive parameterization. LongVideoBench's alternating video-text evaluation suggests that interleaved multimodal inputs pose additional challenges compared to unimodal video tasks, with noticeable impacts on performance, motivating more robust designs for cross-modal agent coordination. Video Recap further shows that hierarchical captioning models retain coherence only across limited temporal abstraction levels, beyond which semantic consistency begins to degrade. Collectively, these findings demonstrate that effective long-duration understanding in training-free systems requires integrated solutions across temporal compression, memory allocation, and hierarchical organization, to maintain reasoning fidelity without incurring prohibitive costs.

3.1.3 Specialized domain understanding

Professional domain evaluation reveals critical knowledge transfer limitations when training-free systems encounter specialized content requiring domain-specific understanding without parameter adaptation. Video-MMMU [93] conducts systematic evaluation across science, engineering, medicine, and other professional domains, identifying notable performance degradation compared to general-domain video tasks when training-free systems process domain-specialized content. These findings underscore key limitations in retrieval-augmented methods that must operate without parameter updates, emphasizing structural constraints for agent-based systems deployed in professional scenarios where specialized terminology and domain-specific reasoning are essential. The benchmark indicates that training-free systems demonstrate moderate accuracy in standard temporal grounding tasks, but face significant challenges in fine-grained localization scenarios that demand precise temporal segmentation, highlighting the inherent difficulty for selection-based methods that lack supervised adaptation. Collectively, these domain-specific evaluations highlight the need for training-free systems to incorporate sophisticated knowledge integration and adaptive temporal reasoning mechanisms to achieve robust performance in specialized domains, while maintaining generalization across diverse content types.

In summary, general-purpose benchmarks establish the foundation by evaluating broad comprehension under frozen parameters. However, they primarily test generalization rather than higher-order cognition, motivating the need for reasoning-oriented benchmarks.

3.2 Reasoning-oriented LVU benchmarks

While general benchmarks establish foundational capabilities for training-free LVU systems across diverse domains and temporal scales, they primarily assess basic comprehension and generalization rather than sophisticated cognitive processes. The inherent limitations of training-free approaches, operating with frozen parameters and relying exclusively on inference-time computation, necessitate specialized evaluation frameworks that probe deeper reasoning capabilities (see Table 3). Unlike supervised systems that can acquire task-specific

Table 3 Overview of video reasoning benchmarks

Benchmark	Year	Reasoning type	Key innovation	Dataset scale	Venue
Temporal and Spatio-Temporal Reasoning					
VideoRefer-Bench	2025	Pixel-level spatial-temporal reasoning	Spatial-temporal understanding and reasoning evaluation	598 videos, 400 description generation annotations, 1 K multi-choice QA pairs	CVPR
V-STaR	2025	Reverse spatio-temporal reasoning	Complete spatio-temporal CoT reasoning thinking chain, RSTR task chain	9 video domains, 2,094 videos	arXiv
TLQA	2025	Temporal logical reasoning	16 temporal logic type definition, comprehensive temporal reasoning evaluation with multiple complexity levels	16 categories of temporal logic, 32 K QA pairs in TLQA-S, 160 K QA pairs in TLQA-L	arXiv
SVBench	2025	Streaming temporal reasoning	Temporal multi-turn question-answering chains, temporal multi-turn dialogue capability evaluation	1.3K streaming videos, 49.9K QA pairs	ICLR
Causal and Physical Reasoning					
CARVE	2025	Causal abductive reasoning	Causal abductive reasoning problem introduction, the first causal abductive reasoning benchmark	2 K videos, over 50 K abductive event pairs	arXiv
PhysBench	2025	Physical world reasoning	Physical world understanding evaluation	4 main categories, 19 subclasses, 10.0K QA pairs	ICLR
MECD+	2025	Multi-event causal reasoning	Multi-event causal discovery task definition, structured event-level causal graph	299 videos	arXiv
ComPhy	2025	Compositional physical reasoning	Diverse simulated scenes and real-world videos, hidden property inference	2 K videos	TPAMI
COVER	2025	Counterfactual video reasoning	Structured query decomposition, fine-grained reasoning evaluation	4 main categories, 13 tasks, around 2.8K videos, 12K-13K QA pairs	ACL Findings
Knowledge-Grounded and Factual Reasoning					
Video SimpleQA	2025	Factual reasoning	External knowledge integration and verification, fact-seeking questions and definitive answers	1,293 videos, 2,030 QA pairs	arXiv
MOMA-QA	2025	Multi-object multi-actor reasoning	Temporal localization and relationship reasoning, frame-level entity-specific annotations	27,586 videos, 300,791 QA pairs	arXiv
CG-Bench	2025	Clue-grounded reasoning	Clue-grouded black/white box evaluation	1,219 videos, 12,129 QA pairs	ICLR
SOK-Bench	2024	Commonsense reasoning	Situated commonsense knowledge integration	10 K videos, 44 K QA pairs,	CVPR
Multi-Modal Compositional Reasoning					
HLV-1K	2025	Long-term reasoning	Diverse reasoning tasks, time-specific queries	1,009 videos, 14,847 QA pairs	ICME
Neptune	2024	Long-term reasoning	Dense and time-aligned captions, broad reasoning skill coverage	2,405 videos, 3,268 QA pairs	arXiv
OVO-Bench	2025	Online Reasoning	Multiple timestamp grounded answering scenarios	644 videos, 2.8K annos	CVPR
VStream-QA	2024	Online Reasoning	Timestamp-based QA pairs, fine-grained long-range video understanding	32 videos, 3.5K QA pairs	arXiv
VideoVista	2024	Comprehensive	Diverse video content and temporal coverage, rich video reasoning tasks	3,402 videos, 25 K QA pairs	arXiv

reasoning patterns through gradient updates, training-free architectures must demonstrate advanced cognitive abilities through orchestrated inference alone, making rigorous reasoning evaluation critical for understanding their operational boundaries and architectural requirements. Reasoning-oriented benchmarks address this gap by systematically evaluating higher-order cognitive capabilities that extend beyond perceptual understanding to encompass spatio-temporal reasoning, causal inference, physical world comprehension, and multi-modal compositional understanding. These benchmarks act as diagnostic tools to determine whether training-free systems perform genuine reasoning or rely on advanced pattern matching, offering critical insights for designing memory-augmented, selection-based, and agent-driven architectures capable of complex temporal reasoning over extended video sequences.

3.2.1 Spatio-temporal reasoning benchmarks

Spatio-temporal reasoning evaluation probes training-free systems' ability to understand complex relationships across both spatial and temporal dimensions, requiring sophisticated inference capabilities that cannot be easily acquired through statistical pattern matching alone. These benchmarks expose fundamental limitations in frozen parameter systems where spatial-temporal understanding must emerge from orchestrated inference rather than learned representations.

Pixel-Level and Object-Centric Spatial-Temporal Assessment: VideoRefer Suite [37] establishes a comprehensive benchmark for pixel-level spatial-temporal object understanding via VideoRefer-700K, a large-scale dataset comprising object-level video instruction samples, with dual evaluation protocols covering both generation and comprehension across multiple temporal reasoning dimensions. The benchmark's multi-agent data engine employs specialized modules for noun extraction, object-level captioning, mask generation, correspondence verification, and summarization, with a particular emphasis on hallucination detection and temporal description consistency. Training-free systems exhibit persistent challenges in maintaining coherent object tracking and relationship understanding across temporal sequences, with notable performance degradation observed when transitioning from single-frame to multi-temporal object reasoning. MOMA-QA [94] targets limitations in temporal and spatial granularity via a Multi-Object Multi-Actor architecture, emphasizing temporal localization, spatial relationship reasoning, and entity-centric querying. It employs ground-truth scene graphs to enable structured representations of objects, relations, and temporal intervals, enhanced by an SGVLM-based predictor for scene-graph-based knowledge modeling. Training-free systems struggle to maintain spatial relationships over time, highlighting limitations in domain-specific spatial reasoning without parameter adaptation.

Explicit Reasoning Chain and Decomposition Assessment: V-STaR [38] introduces the Reverse Spatio-Temporal Reasoning (RSTR) framework, decomposing video understanding into three sequential cognitive components: identifying relevant objects ("what"), determining temporal occurrence ("when"), and analyzing spatial relationships ("where"). The benchmark requires models to perform explicit, human-like multi-step reasoning via coarse-to-fine Chain-of-Thought questions across 14 Video-LLMs, revealing substantial shortcomings in spatio-temporal reasoning when systems must demonstrate stepwise inference rather than rely on pattern matching. Training-free systems exhibit marked performance degradation when required to generate reasoning chains instead of direct answers, reflecting inherent limitations in inference-only reasoning. TimeLogic [39] further explores explicit temporal reasoning through the TLQA framework, generating structured temporal logic questions across 16 distinct categories, focusing on formal logic constructs such as causal, conditional, and temporal dependencies. The benchmark's automated question generation pipeline facilitates large-scale evaluation via TLQA-S and TLQA-L, and findings indicate that current VideoQA models, particularly training-free ones, struggle significantly with formal temporal logic reasoning compared to simpler sequential tasks.

Streaming and Temporal Coherence Assessment: SVBench [40] features temporal multi-turn dialogues for streaming video understanding through consecutive dialogue chains that progress temporally through video content, with temporal dialogue paths representing novel conversation structure mirroring realistic streaming video interaction scenarios. The benchmark's three evaluation modes, SingleQA, ChainQA, and ConQA, provide comprehensive assessment of streaming capabilities, from individual question answering to sustained conversational coherence. Training-free systems face particular challenges in maintaining contextual coherence across extended temporal sequences, revealing fundamental limitations in inference-only temporal memory management.

3.2.2 Causal and physical reasoning benchmarks

Causal and physical reasoning benchmarks evaluate sophisticated cognitive capabilities requiring models to understand underlying mechanisms governing visual phenomena, advancing beyond correlation-based understanding to assess genuine causal inference and physical world comprehension [95, 96]. These benchmarks are particularly challenging for training-free systems that

must infer causal relationships and physical principles through inference alone without gradient-based learning of domain-specific causal patterns.

Causal Discovery and Abductive Reasoning Assessment: CARVE [18] establishes causal abductive reasoning through its Causal Event Relation Network (CERN), which analyzes relationships between video events in both temporal and semantic dimensions, leveraging novel counterfactual synthesis to generate causal relationship labels. The benchmark emphasizes root-cause identification and abductive inference, requiring models to discover foundational causal triggers rather than relying solely on statistical correlations. Training-free systems exhibit clear limitations in causal discovery, particularly in complex inference settings, where their performance substantially lags behind correlation-based methods, underscoring the difficulty of reasoning without causal supervision. MECD+ [19] advances multi-event causal reasoning in long videos through integration of Granger Causality with masked event prediction for fine-grained evaluation. The Event Granger Test compares predicted outcomes with and without premise events to assess causality strength, while Front-Door Adjustment mitigates confounding through formal causal inference techniques. Despite this, training-free systems remain challenged in multi-event causal reasoning, with performance notably lower in complex discovery tasks compared to single-event scenarios across datasets such as ActivityNet, EgoSchema, NExTVideo, and traffic surveillance domains. COVER [20] introduces a comprehensive counterfactual reasoning benchmark spanning abstract-to-concrete and perception-to-cognition axes, comprising 13 tasks distributed across four evaluation quadrants. Its sub-question decomposition structure reveals strong alignment between intermediate reasoning accuracy and overall task performance. Nonetheless, training-free systems exhibit persistent gaps in counterfactual reasoning, performing less reliably on complex hypothetical scenarios than on factual reasoning tasks, highlighting limitations in generalization without causal adaptation.

Physical World Understanding and Property Inference: PhysBench [41] establishes the most comprehensive evaluation framework for physical world understanding, covering object properties, relationships, physical scenes, and physics-driven dynamics across 19 subclasses and 8 capability dimensions. The benchmark's 10,002 entries of interleaved video-image-text data provide systematic assessment spanning thermodynamics, electromagnetism, and optics beyond traditional mechanics. The PhysAgent framework combines VLM generalization with specialized vision model expertise, revealing substantial limitations in current systems across 75 representative Vision-Language Models, with training-free systems showing particular weakness in complex physical phenomena requiring multi-step reasoning compared to basic physical property recognition. ComPhy [97] focuses on inferring non-visible physical properties from motion and interactions, requiring models to understand hidden physical attributes such as mass, electric charge, and friction coefficients. The benchmark's Physical Concept Reasoner (PCR) employs neural-symbolic framework with disentangled components for perception, property learning, dynamics prediction, and symbolic reasoning. The focus on few-shot physical reasoning and compositional understanding exposes key limitations in training-free systems.

3.2.3 Knowledge-grounded and factuality reasoning benchmarks

Knowledge-grounded and factuality reasoning benchmarks evaluate models' ability to integrate external knowledge with video understanding while maintaining factual accuracy and reliability in generated responses. These benchmarks are particularly relevant for training-free systems that cannot acquire new knowledge through parameter updates and must rely on retrieval-augmented inference and knowledge integration mechanisms.

Factuality Assessment and External Knowledge Integration: Video SimpleQA [42] introduces the first comprehensive benchmark specifically designed for factuality evaluation in Large Video Language Models through five-dimensional diagnostic framework requiring integration of domain-specific knowledge, contextual background, and commonsense reasoning while maintaining strict factual accuracy constraints. The benchmark's definitive short-form answer constraint enables reliable evaluation through LLM-as-a-Judge frameworks, while external source verification through credible webpage links ensures answer validation. Training-free systems demonstrate systematic challenges in factual accuracy, with baseline accuracy ranging from 9–80% across different domains, while Retrieval-Augmented Generation (RAG) integration shows consistent F-score improvements (10.3% for Qwen-VL-Max) but reveals computational trade-offs inherent in knowledge integration approaches. Calibration analysis reveals systematic overconfidence across all evaluated models, with Brier score assessments demonstrating misalignment between model confidence and actual accuracy. Training-free systems show particularly severe calibration issues, with confidence scores remaining high (0.7–0.9.7.9) even when accuracy drops to 30–40%, establishing critical implications for real-world deployment where reliable confidence estimation is essential for trustworthy video understanding systems. SOK-Bench [44] contributes situated commonsense reasoning through structured

knowledge integration, employing Situated Knowledge Graphs (SKG) for observable entity extraction and General Knowledge Graphs (GKG) for multi-aspect knowledge generation. The benchmark's Few-Shot Self-Prompting technique enables scalable integration of situated and general knowledge in 12 categories combining objects/actions with spatiotemporal, counterfactual, and purpose reasoning. Human evaluation showing 93.08% QA pair validity validates benchmark quality while revealing significant performance gaps, with GPT-4V achieving only 53.9% accuracy in multiple-choice settings, while training-free systems typically achieve 25–40% accuracy, highlighting substantial limitations in knowledge integration without parameter adaptation.

Long-Form Video Understanding and Clue-Grounded Assessment: CG-Bench [43] introduces clue-grounded evaluation for long-form video understanding, wherein models must retrieve relevant video segments before answering, ensuring that responses reflect genuine comprehension rather than pattern matching or elimination-based heuristics. The benchmark's credibility evaluation framework incorporates both white-box and black-box strategies to assess whether answers are grounded in correct temporal and semantic context. Training-free systems exhibit notable challenges in this setting, particularly in identifying and reasoning over relevant segments, reflecting inherent limitations in long-form inference without adaptation. CG-Bench integrates a comprehensive suite of evaluation metrics, including mIoU, Recall@IoU, Accuracy@IoU, and Clue Retrieval Rate, offering a robust assessment of long-form video reasoning capabilities. The clue-grounded paradigm marks a significant methodological advance by requiring explicit reasoning over retrieved evidence, as opposed to relying on shallow statistical correlations. This evaluation paradigm highlights core weaknesses in current training-free models' ability to perform authentic video understanding when limited to inference-only reasoning mechanisms.

3.2.4 Multi-modal compositional reasoning benchmarks

Multi-modal compositional reasoning benchmarks evaluate the most sophisticated cognitive capabilities, requiring integration of multiple modalities and compositional understanding of complex relationships across diverse temporal scales and domain applications. These benchmarks represent the most challenging evaluation scenarios for training-free systems, requiring orchestrated inference across multiple modalities without specialized multi-modal adaptation.

Long-Form and Hour-Long Video Assessment: HLV-1K [17] introduces hierarchical temporal reasoning across frame-level, within-event, cross-event, and long-horizon tasks through 1,009 hour-long videos with 14,847 QA pairs representing the largest long-video benchmark. The benchmark framework evaluates reasoning across varying temporal granularities, including cross-event integration of distant segments. Training-free systems exhibit marked performance degradation as temporal scale increases, maintaining moderate performance on short-term tasks but struggling significantly with long-range, cross-event inference, highlighting scalability limitations of inference-only models in extended temporal contexts. Neptune [8] proposes a multi-modal compositional reasoning benchmark addressing long-range video understanding with both visual and audio content, spanning summarization, temporal ordering, and intent recognition. The framework introduces the Gemma Equivalence Metric (GEM) for compositional evaluation and includes a scalable pipeline for generating reasoning tasks. Training-free systems exhibit difficulty in integrating multi-modal information, with audio-visual inference performance notably lower than their single-modality counterparts, underscoring limitations in cross-modal generalization without dedicated fusion mechanisms.

Real-Time and Streaming Compositional Assessment: OVO-Bench [98] evaluates temporal compositional reasoning via the Video Chain-of-Time framework, covering backward tracing, real-time perception, and forward-active response strategies. The benchmark includes 12 tasks that require dynamic temporal reasoning in streaming contexts. Forward Active Responding emphasizes delayed decision-making based on accumulated evidence. Training-free systems show constrained performance under these conditions, with noticeably lower accuracy in anticipatory response tasks, reflecting adaptation challenges in dynamic temporal environments. VStream-QA [99] builds on this with a memory-augmented compositional architecture (STAR), enabling fine-grained reasoning under streaming constraints. The dual-process model supports both information retention and low-latency response. VStream-QA provides timestamp-aligned evaluations and reveals that memory-dependent compositional tasks remain challenging for training-free systems, with accuracy levels trailing significantly behind benchmarks established for non-streaming settings.

Comprehensive Multi-Modal Assessment: VideoVista [33] offers a broad compositional benchmark integrating 19 understanding and 8 reasoning tasks, constructed using automated GPT-4o-based pipelines and advanced task analytics. Covering domains such as anomaly detection, interaction analysis, and logical reasoning, the benchmark evaluates both low- and high-level multi-modal skills. Training-free models

consistently underperform across both understanding and compositional reasoning dimensions, reflecting persistent limitations in generalized inference mechanisms in the absence of task-specific adaptation.

These reasoning-oriented benchmarks collectively reveal that training-free systems face substantial limitations in sophisticated cognitive tasks requiring causal inference, physical understanding, knowledge integration, and compositional reasoning. While selection-based methods and memory-augmented frameworks can address some scalability challenges, fundamental constraints remain in systems operating without parameter adaptation for domain-specific reasoning patterns, establishing critical design requirements for future training-free architectures capable of advanced temporal reasoning capabilities.

Beyond reasoning, egocentric benchmarks introduce unique embodied and first-person challenges, which we turn to next.

3.3 Egocentric LVU benchmarks

Egocentric benchmarks can be further divided into three thematic subgroups: (1) long-term reasoning & causal inference, (2) scene-text reasoning & alignment, and (3) multi-dimensional evaluation.

The comprehensive evaluation frameworks examined thus far have primarily addressed training-free LVU systems through general comprehension, reasoning-oriented assessment, and scalability analysis across third-person video content. However, egocentric video understanding introduces fundamentally distinct challenges that extend beyond conventional temporal reasoning to encompass embodied cognition, spatial-temporal grounding from first-person perspectives, and dynamic environmental interaction. These unique characteristics necessitate specialized evaluation protocols that assess training-free systems' capability to understand human-centered visual experiences while maintaining computational efficiency and generalization across diverse egocentric scenarios (see Table 4). Egocentric videos, captured from wearable devices positioned at human eye-level, present systematic challenges including frequent camera motion, close-range object interactions, partial observability constraints, and substantial visual appearance variability due to personalized viewpoints [45, 46]. These characteristics fundamentally alter the computational requirements for training-free systems, which must demonstrate robust performance across dynamic viewpoints without parameter adaptation while maintaining temporal coherence across extended first-person sequences. Unlike third-person videos where camera positions remain relatively stable, egocentric content requires sophisticated spatial reasoning capabilities that can adapt to continuously changing reference frames and partial occlusions inherent in human-centered recording scenarios. The evaluation of training-free systems within egocentric contexts reveals critical limitations in current architectures, particularly regarding long-range temporal dependency modeling, embodied spatial reasoning, and multi-modal integration across visual and textual information sources. These benchmarks provide essential diagnostic frameworks for probing the limits of inference-only approaches under the cognitive demands of first-person video understanding, offering insights toward advancing training-free architectures for embodied intelligence.

Table 4 Overview of egocentric video understanding benchmarks

Benchmark	Year	Task category	Video length	Dataset scale	Key features	Venue
Long-Term Temporal Reasoning & Causal Inference						
EgoSchema	2023	Long-Horizon QA	Avg. 180 sec	5,063 videos, 5,063 QA pairs	Long-form egocentric video understanding	NeurIPS
EgoTaskQA	2022	Task-Centric QA	-	40 K QA pairs	Spatial, temporal, and causal long-form video understanding	NeurIPS
Ego4D-HCap	2024	Long-Horizon QA	Avg. 1708 sec (long-range video subset)	5.27M short-range videos, 17.5K medium-length videos, 8.3K long-range videos	Egocentric videos with hierarchical length and captions	CVPR
EgoLifeQA	2025	Long-Horizon QA	Avg. 44.3 hr	6 extreme long videos, 6 K QA pairs	Extreme long-form egocentric video understanding	CVPR
Scene Text Reasoning & Visual-Text Alignment						
EgoTextVQA	2025	Text-Aware QA	Avg. 102 sec	1,507 videos, 7,064 QA pairs	Scene text integrated egocentric video understanding	CVPR
Multi-Dimensional Egocentric Evaluation						
EgoThink	2024	Multi-Skill QA	-	700 QA pairs	diverse evaluation tasks, egocentric spatial awareness	CVPR

3.3.1 Long-term temporal reasoning & causal inference

Egocentric temporal reasoning evaluation reveals fundamental limitations in training-free systems' ability to maintain coherent understanding across extended first-person sequences where traditional temporal cues are disrupted by dynamic viewpoints and partial observability. These benchmarks establish critical performance boundaries for inference-only architectures operating under the unique constraints of human-centered video content, particularly when complex causal relationships must be inferred from visual evidence alone.

Long-Term Coherence and Action Understanding: EgoSchema [45] establishes a rigorous benchmark for assessing long-form egocentric video understanding, encompassing 5,000 multiple-choice questions derived from 250 hours of Ego4D content. Each query necessitates the integration of temporal information across 3-minute video segments. The core innovation of this benchmark lies in the introduction of a temporal certificate length metric, which quantifies the minimum video segment duration required for reliable human comprehension. This metric reveals that the median certificate length for egocentric video understanding is 100 seconds, in stark contrast to the 18 seconds typically observed in third-person benchmarks. This substantial 5.7-fold increase in temporal dependency uncovers significant limitations in training-free systems, which exhibit a sharp decline in performance from 65–70% accuracy on 30-second video segments to a mere 25–33% on extended egocentric sequences. The benchmark underscores that current training-free architectures are incapable of preserving semantic coherence across the long temporal horizons essential for egocentric understanding, highlighting the profound challenges posed by inference-only models when faced with continuously evolving first-person perspectives. Ego4D-HCap [100] introduces a hierarchical video dataset tailored for evaluating long-range temporal reasoning and causal reasoning in egocentric video understanding. Spanning video segments of varying lengths (with an average of 1708 seconds), the dataset comprises 5.27 million short-range videos, 17.5K medium-length videos, and 8.3K long-range videos. It facilitates in-depth analysis of temporal reasoning over extended video sequences, accompanied by hierarchical captions, thus addressing the complexities inherent in long-range video understanding from an egocentric perspective. EgoLifeQA [101] further advances the field by focusing on extreme long-form egocentric video understanding. Featuring six videos, each averaging 44.3 hours in length, and 6,000 question-answer pairs, this benchmark tests systems on their ability to perform spatial, temporal, and causal reasoning over extended durations. It challenges models to maintain semantic coherence and integrate information across vast time spans, pushing the boundaries of current egocentric video understanding.

Hierarchical Task Understanding and Causal Modeling: EgoTaskQA [46] advances causal inference evaluation through systematic assessment of causal reasoning within egocentric task hierarchies, employing 40,000 question-answer pairs across descriptive, predictive, explanatory, and counterfactual reasoning categories. The benchmark's integration of LEMMA dataset annotations enables precise evaluation of object state transition understanding and action dependency modeling, revealing systematic limitations in training-free systems' ability to infer causal relationships from first-person visual evidence. State-of-the-art video reasoning models (e.g., HCRN, ClipBERT) achieve only around 40% accuracy on multi-agent and counterfactual reasoning, while human performance reaches from 74% to 88% across different reasoning categories. These findings highlight that inference-only systems struggle with the complex temporal reasoning patterns required for understanding hierarchical task structures from egocentric perspectives, revealing critical limitations in systems operating without parameter adaptation for domain-specific causal patterns.

3.3.2 Scene text reasoning & visual-text alignment

The integration of textual information within egocentric video understanding presents unique challenges for training-free systems, requiring sophisticated multi-modal fusion capabilities that can maintain consistent text-visual alignment across dynamic first-person viewpoints without parameter adaptation [102]. This evaluation domain exposes critical limitations in inference-only architectures when confronted with the complex multi-modal integration demands characteristic of real-world egocentric scenarios. EgoTextVQA [47] introduces systematic evaluation of scene text reasoning within egocentric contexts through 1,500 videos and 7,000 questions spanning outdoor driving and indoor household scenarios. The benchmark requires models to integrate dynamic textual information, including road signs, product labels, and environmental text, with visual context to generate situation-appropriate responses, revealing severe limitations in training-free systems' multi-modal integration capabilities. State-of-the-art models achieve only 33% accuracy compared to human performance ranging from 27% to 43% across different scenarios, highlighting fundamental deficits in dynamic text localization and cross-frame semantic integration when operating without parameter adaptation. The benchmark exposes particular weaknesses in training-free systems' ability to maintain consistent text-visual alignment across changing viewpoints and temporal sequences, establishing critical

constraints for inference-only architectures in real-world multi-modal scenarios where textual information must be dynamically tracked and semantically integrated with evolving visual contexts.

3.3.3 Multi-dimensional egocentric evaluation

Comprehensive cognitive capability assessment across multiple dimensions reveals systematic limitations in training-free systems when confronted with the diverse cognitive demands of egocentric understanding. This evaluation framework exposes fundamental constraints in inference-only architectures across the broad spectrum of capabilities required for robust first-person video comprehension. EgoThink [48] provides systematic evaluation across six core cognitive dimensions, object recognition, activity understanding, spatial navigation, temporal reasoning, causal inference, and intention prediction, through 700 carefully annotated samples spanning 12 fine-grained assessment categories. Despite leveraging state-of-the-art vision-language models, experimental results show that even the strongest model (GPT-4V) achieves only around 60–65% accuracy on average, while most open-source models remain at 20–40% across different dimensions. Particularly in planning and sequential reasoning tasks, models exhibit inconsistent performance and fail to maintain stable egocentric spatial awareness over temporal sequences, highlighting their inability to coherently aggregate sequential visual cues for decision-making. Although some dimensions (e.g., localization and planning) yield slightly higher scores compared to others, the benchmark exposes a fundamental gap between current training-free architectures and the comprehensive reasoning capabilities required for robust first-person understanding. These findings underscore that inference-only systems still struggle to emulate the integrated, self-centered reasoning patterns necessary for human-level egocentric cognition.

The comprehensive benchmark analysis across general comprehension, reasoning-oriented assessment, and egocentric understanding domains establishes a systematic diagnostic framework for evaluating training-free LVU systems, revealing consistent limitations that fundamentally constrain inference-only architectures when applied to sophisticated video understanding tasks. While training-free approaches demonstrate reasonable performance on basic temporal understanding and content recognition tasks, they face severe degradation when confronted with complex reasoning demands (accuracy drops of 40–55% compared to direct comprehension), extended temporal dependencies (25–33% accuracy on sequences exceeding 30 seconds), and embodied interaction scenarios (75% performance deficit compared to human capabilities). These systematic limitations highlight the fundamental trade-offs inherent in training-free paradigms: while they offer significant advantages in computational efficiency, deployment flexibility, and cross-domain generalization, they cannot match the specialized reasoning capabilities achieved through parameter adaptation. The convergence of these findings across diverse evaluation domains establishes critical design constraints for future training-free architectures, indicating that advances in inference-time coordination, memory-augmented frameworks, and agent-based reasoning strategies will be essential for bridging the performance gap between training-free systems and specialized approaches while preserving the core advantages that define this paradigm.

4 Challenges

4.1 Practical deployment value vs. challenges

The comprehensive evaluation of training-free LVU systems across diverse benchmarks reveals fundamental challenges that limit their effectiveness in real-world applications. These challenges span computational constraints, architectural limitations, and domain-specific complexities that collectively define the operational boundaries of current approaches.

Computational and Memory Bottlenecks: Training-free LVU systems face severe computational constraints when processing extended video sequences. Frame-level feature extraction across thousands of frames results in computational overload that quickly exceeds hardware constraints, while most existing architectures impose fixed context windows that fragment video continuity, making it difficult to capture long-range dependencies and hierarchical temporal structure. The problem of visual-token redundancy becomes particularly pronounced in long video settings, where seamlessly integrated compression methods must perform fine-grained token reduction via content-aware or multi-stage mechanisms to remain computationally tractable. The challenge extends beyond raw computational demands to memory management. Memory-augmented approaches focus on boosting long-video reasoning capacity and speed under limited computational budgets, but real-time streaming applications require joint algorithmic and architectural optimization to achieve low latency and high throughput. Even with sophisticated compression techniques, maintaining semantic fidelity while reducing token complexity remains a persistent challenge.

Temporal Reasoning and Long-Range Dependencies: All models achieve significantly higher scores on short tasks than medium and long tasks, indicating that LVU remains a major challenge. For example, GPT4o (32 frames) scores 71.4 on short tasks but only 55.2 on long tasks (a 16.2-point gap). This systematic performance

degradation reveals fundamental limitations in temporal reasoning capabilities. Training-free systems exhibit marked performance degradation as the temporal scale increases, maintaining moderate performance on short-term tasks but struggling significantly with long-range, cross-event inference, highlighting scalability limitations of inference-only models in extended temporal contexts. The challenge is particularly acute in egocentric scenarios, where training-free systems demonstrate catastrophic performance degradation on 30-second segments to extended egocentric sequences.

Complex Reasoning and Causal Inference: Beyond basic comprehension, LVU systems must perform sophisticated cognitive tasks that expose severe limitations in training-free approaches. Training-free systems exhibit clear limitations in causal discovery, particularly in complex inference settings, where their performance substantially lags behind correlation-based methods. Training-free systems remain challenged in multi-event causal reasoning, with performance notably lower in complex discovery tasks compared to single-event scenarios. Training-free systems exhibit persistent gaps in counterfactual reasoning, performing less reliably on complex hypothetical scenarios than on factual reasoning tasks, highlighting limitations in generalization without causal adaptation. These reasoning challenges are compounded by the need for explicit reasoning chains, where training-free systems exhibit marked performance degradation when required to generate reasoning chains instead of direct answers, reflecting inherent limitations in inference-only reasoning.

Multi-Modal Integration and Alignment: The integration of multiple modalities presents significant challenges for training-free systems. Training-free systems exhibit difficulty in integrating multi-modal information, with audio-visual inference performance notably lower than their single-modality counterparts, underscoring limitations in cross-modal generalization without dedicated fusion mechanisms. This challenge is particularly evident in egocentric scenarios with textual content. State-of-the-art models achieve low accuracy compared to human performance across different scenarios, highlighting fundamental deficits in dynamic text localization and cross-frame semantic integration when operating without parameter adaptation.

Domain Generalization and Knowledge Grounding: Training-free systems achieve consistent performance only within domains similar to their pre-training distribution, establishing fundamental constraints for agent-based reasoning systems. Video-MMMU conducts systematic evaluation across science, engineering, medicine, and other professional domains, identifying notable performance degradation compared to general-domain video tasks when training-free systems process domain-specialized content. The challenge extends to factual accuracy and knowledge integration. Calibration analysis reveals systematic overconfidence across all evaluated models, with confidence scores remaining high (0.7–0.9.7.9) even when accuracy drops to 30–40%.

Egocentric and Embodied Understanding: Egocentric video understanding presents unique challenges that significantly strain the capabilities of training-free systems. Unlike third-person videos with relatively stable viewpoints, egocentric content features constant camera motion, close-range interactions, partial observability, and high appearance variability. These factors demand robust spatial and temporal reasoning without the benefit of parameter adaptation. Training-free systems, which rely solely on inference, often falter in this setting. Maintaining temporal coherence across long, dynamic sequences is difficult due to shifting perspectives and disrupted visual continuity. Spatial reasoning is also challenged by frequent occlusions and unstable reference frames inherent in first-person perspectives. Moreover, egocentric understanding often involves interpreting hierarchical tasks and causal relationships in visually complex, personalized environments. Without specialized adaptation, current training-free architectures struggle to integrate multi-modal cues and model long-range dependencies, revealing limits in training-free methods to first-person video comprehension.

Agent Coordination and Memory Management: While agent-based approaches show promise, they face significant coordination challenges. MLVU highlights limitations of current architectures during task switching within continuous video contexts, revealing performance degradation when coordinating multiple cognitive processes without adaptive parameterization. Adaptive hierarchical structures must strike a balance between information extraction and contextual coherence, while memory and compression techniques must reduce dependency on large-scale training and resources. These challenges collectively define the current limitations of training-free LVU systems and establish the operational constraints within which these approaches must function. Addressing these challenges requires not merely incremental improvements but fundamental advances in architectural design, inference-time computation, and multi-modal integration strategies.

4.2 Future research directions

Hybrid Paradigms: Lightweight, parameter-efficient adaptation strategies (e.g., LoRA, modular adapters, or mixture-of-experts layers) could serve as a complementary layer to training-free orchestration, enabling systems to selectively refine critical components while preserving

frozen backbones. Such approaches [50, 103–105] offer a middle ground between fully static inference and costly retraining, helping models sustain performance under increasing video length and modality complexity, while also enhancing adaptability to domain-specific scenarios and emerging tasks.

Adaptive Memory and Compression: Developing more interpretable and resource-aware memory architectures [106, 107], in conjunction with multi-granularity compression techniques, can help alleviate the exponential growth in computation while maintaining semantic fidelity across extended sequences. To ensure practical scalability, future benchmarks should go beyond accuracy and explicitly incorporate sustainability metrics such as inference latency, token utilization efficiency, and long-horizon robustness, providing a more holistic assessment of system viability.

Cross-Modal Orchestration: Equally important is the design of dynamic retrieval and reasoning strategies that can explicitly capture and coordinate dependencies across modalities (e.g., text, audio, and vision). Such cross-modal orchestration is crucial for moving beyond vision-centric pipelines, allowing training-free systems to generalize effectively to richer, real-world scenarios where multiple modalities interact in complex and temporally extended ways.

4.3 Discussion

Because training-free LVU systems emphasize reasoning rather than parameter learning, their modular orchestration—through keyframe selection, memory structures, and agent-based Chain-of-Thought prompting—naturally exposes intermediate artifacts such as selected frames, memory traces, and reasoning chains. These artifacts make the inference process more transparent and provide diagnostic signals that can be inspected by users and developers. In contrast to end-to-end trained models, such modularity enables failures to be localized to specific stages (e.g., sampling, retrieval, or reasoning), supporting systematic debugging, error attribution, and adaptive refinement without retraining.

To make these reasoning-centered approaches reliable in applied contexts, future work should focus on strengthening interpretability and trust. Promising directions include verifiable reasoning traces grounded in retrieved evidence, calibration techniques that align confidence with accuracy, and human-in-the-loop protocols that render outputs auditable in high-stakes decision-making. In addition, benchmarks should evolve beyond accuracy alone to incorporate explainability and robustness criteria—such as interpretability, factual grounding, and resilience to noisy or adversarial inputs. Embedding these dimensions will not only make the survey of training-free LVU more forward-looking but also ensure that such systems remain both scalable and trustworthy for real-world deployment.

5 Conclusion

This survey positions training-free LVU as a paradigm shift in video analysis, demonstrating that frozen models, when orchestrated at inference time, can perform complex temporal reasoning without task-specific optimization. Despite 40–55% performance drops in reasoning tasks and up to 75% in embodied settings, these systems offer advantages in efficiency, generalization, and scalability. The combination of selection, memory, and agent-based strategies redefines how systems approach video understanding, replacing heavy optimization with cognitively-inspired inference. While current models face limits in coherence, planning, and multimodal reasoning, they provide a flexible foundation for scalable deployment across varied domains. As video content proliferates across real-world applications, training-free approaches offer a sustainable path forward. Their success depends on principled architectural innovation and rigorous evaluation. By addressing current limitations while preserving core strengths, training-free LVU systems hold promise for achieving robust, human-level temporal understanding through efficient, generalizable inference.

Abbreviations

LVU	Long Video Understanding
VLM	Vision-Language Model
MLLM	Multimodal Large Language Model
LLM	Large Language Model
CoT	Chain-of-Thought
RAG	Retrieval-Augmented Generation
QA	Question Answering
OCR	Optical Character Recognition
KV	Key-Value
ASS	Average Similarity Score
GCL	Group Contrastive Learning
NIF	Necessary Information Frames
PFR	Pivot Frame Retrieval
RSTR	Reverse Spatio-Temporal Reasoning
CERN	Causal Event Relation Network
PCR	Physical Concept Reasoner
SKG	Situated Knowledge Graphs
GKG	General Knowledge Graphs
GEM	Gemma Equivalence Metric
IoU	Intersection over Union
mIoU	mean Intersection over Union
FLOPs	Floating Point Operations
TSC	Temporal Scene Clustering
CKD	Coarse Keyframe Detector
FKD	Fine Keyframe Detector

Acknowledgements

We would like to thank the researchers in the field of training-free long video understanding for their foundational work. We are also grateful to the creators of the open-source benchmarks, which have greatly supported our research.

Authors' contributions
This manuscript represents a collaborative effort among all authors. Jingren Liu took the lead in drafting the manuscript and conducting the core analyses. Yun Wang and Long Zhang were primarily responsible for the development of methodologies and execution of the experiments. Yiheng Wang, Shuning Xu, Ling Wang, and Jiaqi Yan contributed substantially to the benchmarking studies and identification of open challenges. Dell Zhang provided critical feedback and substantive revisions that significantly enhanced the manuscript's clarity and rigor. Xiangyu Chen served as the project administrator. All authors contributed to manuscript revisions and approved the final version.

Funding
Not applicable.

Data availability
All data used in this manuscript is publicly available. No new data was used for the research described in the manuscript.

Code availability
The codes of baseline methods are from the public website.

Declarations

Competing interests
The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work in this manuscript.

Author details
[1]Institute of Artificial Intelligence (TeleAI), China Telecom, Shanghai, China. [2]Tianjin University, Tianjin, China. [3]City University of Hong Kong, Hong Kong, China. [4]University of Science and Technology of China, Hefei, China. [5]University of Macau, Macau, China. [6]Nanjing University, Nanjing, China.

Received: 31 July 2025 Revised: 15 September 2025 Accepted: 19 September 2025
Published online: 09 October 2025

References

1. L. Zhang, P. Song, Z. Duan, S. Wang, X. Chang, X. Yang, Video corpus moment retrieval with query-specific context learning and progressive localization. IEEE Trans. Circuits Syst. Video Technol. **35**, (6)5659-5670. (2025) https://doi.org/10.1109/TCSVT.2025.3530570
2. P. Song, L. Zhang, L. Lan, W. Chen, D. Guo, X. Yang, M. Wang, Towards efficient partially relevant video retrieval with active moment discovering. (2025). arXiv preprint arXiv:2504.10920. https://arxiv.org/abs/2504.10920
3. L. Zhang, P. Song, J. Dong, K. Li, X. Yang, Enhancing partially relevant video retrieval with robust alignment learning. (2025). arXiv preprint arXiv:2509.01383. https://arxiv.org/abs/2509.01383
4. H. An, W. Hu, S. Huang, S. Huang, R. Li, Y. Liang, J. Shao, Y. Song, Z. Wang, C. Yuan et al., Ai flow: Perspectives, scenarios, and approaches. (2025). arXiv preprint arXiv:2506.12479. https://arxiv.org/abs/2506.12479
5. D. Zhang, X. Chen, J. Luo, M. Jia, C. Sun, R. Ren, J. Liu, H. Sun, X. Li, Infinite video understanding. (2025). arXiv preprint arXiv:2507.09068. https://arxiv.org/abs/2507.09068
6. H. Wu, D. Li, B. Chen, J. Li, Longvideobench: a benchmark for long-context interleaved video-language understanding. Adv. Neural. Inf. Process. Syst. **37**, 28828–28857 (2024)
7. J. Zhou, Y. Shu, B. Zhao, B. Wu, S. Xiao, X. Yang, Y. Xiong, B. Zhang, T. Huang, Z. Liu, Mlvu: A comprehensive benchmark for multi-task long video understanding. (2024). arXiv preprint arXiv:2406.04264. https://arxiv.org/abs/2406.04264
8. A. Nagrani, M. Zhang, R. Mehran, R. Hornung, N.B. Gundavarapu, N. Jha, A. Myers, X. Zhou, B. Gong, C. Schmid et al., Neptune: The long orbit to benchmarking long video understanding. (2024). arXiv preprint arXiv:2412.09582. https://arxiv.org/abs/2412.09582
9. H. Zhang, Y. Xu, S. Huang, X. Li, Data augmentation of contrastive learning is estimating positive-incentive noise. (2024). arXiv preprint arXiv:2408.09929. https://arxiv.org/abs/2408.09929
10. H. Zhang, S. Huang, Y. Guo, X. Li, Variational positive-incentive noise: how noise benefits models. IEEE Trans. Pattern Anal. Mach. Intell. **47**,(9) 8313–8320. (2025) https://doi.org/10.1109/TPAMI.2025.3575295
11. X. Li, Positive-incentive noise. IEEE Trans. Neural Netw. Learn. Syst. **35**(6), 8708–8714 (2022)
12. S. Huang, Y. Xu, H. Zhang, X. Li, Learn beneficial noise as graph augmentation. (2025). arXiv preprint arXiv:2505.19024. https://arxiv.org/abs/2505.19024
13. Y. Zhang, Z. Zhao, Z. Chen, Z. Ding, X. Yang, Y. Sun, Beyond training: Dynamic token merging for zero-shot video understanding. (2024). arXiv preprint arXiv:2411.14401. https://arxiv.org/abs/2411.14401
14. M. Wang, S. Chen, K. Kersting, V. Tresp, Y. Ma, Metok: Multi-stage event-based token compression for efficient long video understanding. (2025). arXiv preprint arXiv:2506.02850. https://arxiv.org/abs/2506.02850
15. Y. Guo, W. Dong, J. Song, S. Zhu, X. Zhang, H. Yang, Y. Wang, Y. Du, X. Chen, B. Zheng, Fila-video: Spatio-temporal compression for fine-grained long video understanding. (2025). arXiv preprint arXiv:2504.20384. https://arxiv.org/abs/2504.20384
16. Y. Wang, K. Li, L. Wang, J. Hu, D.O. Wu, Y. Guo, Adstereo: efficient stereo matching with adaptive downsampling and disparity alignment. IEEE Trans. Image Process. **34**, 1204–1218. (2025) https://doi.org/10.1109/TIP.2025.3540282
17. H. Zou, T. Luo, G. Xie, V.X.J. Zhang, F. Lv, G. Wang, J. Chen, Z. Wang, H. Zhang, H. Zhang, Hlv-1k: A large-scale hour-long video benchmark for time-specific long video understanding. (2025). arXiv preprint arXiv:2501.01645. https://arxiv.org/abs/2501.01645
18. T.M. Le, V. Le, K. Do, S. Gupta, S. Venkatesh, T. Tran, Finding the trigger: Causal abductive reasoning on video events. (2025). arXiv preprint arXiv:2501.09304. https://arxiv.org/abs/2501.09304
19. T. Chen, H. Liu, Y. Wang, Y. Chen, T. He, C. Gan, H. He, W. Lin, Mecd+: Unlocking event-level causal graph discovery for video reasoning. (2025). arXiv preprint arXiv:2501.07227. https://arxiv.org/abs/2501.07227
20. Q. Zhou, Y. Gong, G. Bao, H. Qiu, J. Li, X. Zhu, H. Zhang, Y. Zhang, Reasoning is all you need for video generalization: A counterfactual benchmark with sub-question evaluation. (2025). arXiv preprint arXiv:2503.10691. https://arxiv.org/abs/2503.10691
21. X. Tang, J. Qiu, L. Xie, Y. Tian, J. Jiao, Q. Ye, Adaptive keyframe sampling for long video understanding. in *Proceedings of the Computer Vision and Pattern Recognition Conference* (CVPR). (Computer Vision Foundation / IEEE, Nashville, TN, USA 2025). pp. 29118–29128.
22. K. Hu, F. Gao, X. Nie, P. Zhou, S. Tran, T. Neiman, L. Wang, M. Shah, R. Hamid, B. Yin et al., M-llm based video frame selection for efficient video understanding. in *Proceedings of the Computer Vision and Pattern Recognition Conference* (CVPR). (Computer Vision Foundation / IEEE, Nashville, TN, USA 2025). pp. 13702–13712
23. D. Cheng, M. Li, J. Liu, Y. Guo, B. Jiang, Q. Liu, X. Chen, B. Zhao, Enhancing long video understanding via hierarchical event-based memory. (2024). arXiv preprint arXiv:2409.06299. https://arxiv.org/abs/2409.06299
24. Z. Wang, S. Yu, E. Stengel-Eskin, J. Yoon, F. Cheng, G. Bertasius, M. Bansal, Videotree: Adaptive tree-based video representation for llm reasoning on long videos. in *Proceedings of the Computer Vision and Pattern Recognition Conference* (CVPR). (Nashville, TN, USA. 2025). pp. 3272–3283
25. S. Santos, A. Farinhas, D.C. McNamee, A.F. Martins, ∞-Video: A training-free approach to long video understanding via continuous-time memory consolidation. (2025). arXiv preprint arXiv:2501.19098. https://arxiv.org/abs/2501.19098
26. B. Chen, Z. Yue, S. Chen, Z. Wang, Y. Liu, P. Li, Y. Wang, Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. (2025). arXiv preprint arXiv:2503.10200. https://arxiv.org/abs/2503.10200
27. Z. Zhi, Q. Wu, W. Li, Y. Li, K. Shao, K. Zhou et al., Videoagent2: Enhancing the llm-based agent system for long-form video understanding by uncertainty-aware cot. (2025). arXiv preprint arXiv:2504.04471. https://arxiv.org/abs/2504.04471

28. X. Jiang, Y. Pang, X. Li, J. Pan, Y. Xie, Deep neural networks with elastic rectified linear units for object recognition. Neurocomputing **275**, 1132–1139 (2018)
29. D. Tao, X. Li, S.J. Maybank, X. Wu, Human carrying status in visual surveillance. in *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*. (IEEE, New York, NY, USA 2006). **2**, pp. 1670–1677
30. J. Han, D. Zhang, S. Wen, L. Guo, T. Liu, X. Li, Two-stage learning to predict human eye fixations via sdaes. IEEE Trans. Cybern. **46**(2), 487–498 (2015)
31. K. Li, Y. Wang, Y. He, Y. Li, Y. Wang, Y. Liu, Z. Wang, J. Xu, G. Chen, P. Luo et al., Mvbench: A comprehensive multi-modal video understanding benchmark. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR). (Computer Vision Foundation / IEEE, Seattle, WA, USA 2024). pp. 22195–22206
32. V. Patraucean, L. Smaira, A. Gupta, A. Recasens, L. Markeeva, D. Banarse, S. Koppula, M. Malinowski, Y. Yang, C. Doersch et al., Perception test: a diagnostic benchmark for multimodal video models. Adv. Neural. Inf. Process. Syst. **36**, 42748–42761 (2023)
33. Y. Li, X. Chen, B. Hu, L. Wang, H. Shi, M. Zhang, Videovista: A versatile benchmark for video understanding and reasoning. (2024). arXiv preprint arXiv:2406.11303. https://arxiv.org/abs/2406.11303
34. C. Fu, Y. Dai, Y. Luo, L. Li, S. Ren, R. Zhang, Z. Wang, C. Zhou, Y. Shen, M. Zhang et al., Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. in *Proceedings of the Computer Vision and Pattern Recognition Conference* (CVPR). (Computer Vision Foundation / IEEE, Nashville, TN, USA 2025). pp. 24108–24118
35. E. Song, W. Chai, G. Wang, Y. Zhang, H. Zhou, F. Wu, H. Chi, X. Guo, T. Ye, Y. Zhang et al., Moviechat: From dense token to sparse memory for long video understanding. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR). (Computer Vision Foundation / IEEE, Seattle, WA, USA 2024). pp. 18221–18232
36. W. Wang, Z. He, W. Hong, Y. Cheng, X. Zhang, J. Qi, X. Gu, S. Huang, B. Xu, Y. Dong et al., Lvbench: An extreme long video understanding benchmark. (2024). arXiv preprint arXiv:2406.08035. https://arxiv.org/abs/2406.08035
37. Y. Yuan, H. Zhang, W. Li, Z. Cheng, B. Zhang, L. Li, X. Li, D. Zhao, W. Zhang, Y. Zhuang et al., Videorefer suite: Advancing spatial-temporal object understanding with video llm. in *Proceedings of the Computer Vision and Pattern Recognition Conference* (CVPR). (Computer Vision Foundation / IEEE, Nashville, TN, USA 2025). pp. 18970–18980
38. Z. Cheng, J. Hu, Z. Liu, C. Si, W. Li, S. Gong, V-star: Benchmarking video-llms on video spatio-temporal reasoning. (2025). arXiv preprint arXiv:2503.11495. https://arxiv.org/abs/2503.11495
39. S. Swetha, H. Kuehne, M. Shah, Timelogic: A temporal logic benchmark for video qa. (2025). arXiv preprint arXiv:2501.07214. https://arxiv.org/abs/2501.07214
40. Z. Yang, Y. Hu, Z. Du, D. Xue, S. Qian, J. Wu, F. Yang, W. Dong, C. Xu, Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. (2025). arXiv preprint arXiv:2502.10810. https://arxiv.org/abs/2502.10810
41. W. Chow, J. Mao, B. Li, D. Seita, V. Guizilini, Y. Wang, Physbench: Benchmarking and enhancing vision-language models for physical world understanding. (2025). arXiv preprint arXiv:2501.16411. https://arxiv.org/abs/2501.16411
42. M. Cao, P. Hu, Y. Wang, J. Gu, H. Tang, H. Zhao, J. Dong, W. Yu, G. Zhang, I. Reid et al., Video simpleqa: Towards factuality evaluation in large video language models. (2025). arXiv preprint arXiv:2503.18923. https://arxiv.org/abs/2503.18923
43. G. Chen, Y. Liu, Y. Huang, Y. He, B. Pei, J. Xu, Y. Wang, T. Lu, L. Wang, Cg-bench: Clue-grounded question answering benchmark for long video understanding. (2024). arXiv preprint arXiv:2412.12075. https://arxiv.org/abs/2412.12075
44. A. Wang, B. Wu, S. Chen, Z. Chen, H. Guan, W.N. Lee, L.E. Li, C. Gan, in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. Sok-bench: A situated video reasoning benchmark with aligned open-world knowledge (CVPR). (Computer Vision Foundation / IEEE, Seattle, WA, USA 2024). pp. 13384–13394
45. K. Mangalam, R. Akshulakov, J. Malik, Egoschema: a diagnostic benchmark for very long-form video language understanding. Adv. Neural. Inf. Process. Syst. **36**, 46212–46244 (2023)
46. B. Jia, T. Lei, S.C. Zhu, S. Huang, Egotaskqa: understanding human tasks in egocentric videos. Adv. Neural. Inf. Process. Syst. **35**, 3343–3360 (2022)
47. S. Zhou, J. Xiao, Q. Li, Y. Li, X. Yang, D. Guo, M. Wang, T.S. Chua, A. Yao, Egotextvqa: Towards egocentric scene-text aware video question answering. in *Proceedings of the Computer Vision and Pattern Recognition Conference* (CVPR). (Computer Vision Foundation / IEEE, Nashville, TN, USA: 2025). pp. 3363–3373
48. S. Cheng, Z. Guo, J. Wu, K. Fang, P. Li, H. Liu, Y. Liu, Egothink: Evaluating first-person perspective thinking capability of vision-language models. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR). (Computer Vision Foundation / IEEE, Seattle, WA, USA 2024). pp. 14291–14302
49. Z. Cheng, Y. Tu, R. Li, S. Dai, J. Hu, S. Hu, J. Li, Y. Shi, T. Yu, W. Chen et al., Embodiedeval: Evaluate multimodal llms as embodied agents. (2025). arXiv preprint arXiv:2501.11858. https://arxiv.org/abs/2501.11858
50. A. Ma, J. Feng, K. Cao, J. Wang, Y. Wang, Q. Zhang, Z. Zhang, Lay2story: Extending diffusion transformers for layout-togglable story generation. (2025). arXiv preprint arXiv:2508.08949. https://arxiv.org/abs/2508.08949
51. X. Liu, Z. Wang, Y. Han, Y. Wang, J. Yuan, J. Song, B. Zheng, L. Zhang, S. Huang, H. Chen, Global compression commander: Plug-and-play inference acceleration for high-resolution large vision-language models. (2025). arXiv preprint arXiv:2501.05179. https://arxiv.org/abs/2501.05179
52. H. Yuan, Z. Liu, M. Qin, H. Qian, Y. Shu, Z. Dou, J.R. Wen, Memory-enhanced retrieval augmentation for long video understanding. (2025). arXiv preprint arXiv:2503.09149.
53. Z. Ning, G. Liu, Q. Jin, W. Ding, M. Guo, J. Zhao, Livevlm: Efficient online video understanding via streaming-oriented kv cache and retrieval. (2025). arXiv preprint arXiv:2505.15269. https://arxiv.org/abs/2505.15269
54. B. Schneider, D. Jiang, C. Du, T. Pang, W. Chen, Quickvideo: Real-time long video understanding with system algorithm co-design. (2025). arXiv preprint arXiv:2505.16175. https://arxiv.org/abs/2505.16175
55. X. Tan, Y. Ye, Y. Luo, Q. Wan, F. Liu, Z. Cai, Rag-adapter: A plug-and-play rag-enhanced framework for long video understanding. (2025). arXiv preprint arXiv:2503.08576. https://arxiv.org/abs/2503.08576
56. X. Wang, Q. Si, J. Wu, S. Zhu, L. Cao, L. Nie, Retake: Reducing temporal and knowledge redundancy for long video understanding. (2024). arXiv preprint arXiv:2412.20504. https://arxiv.org/abs/2412.20504
57. X. Liu, Y. Wang, J. Ma, L. Zhang, Video compression commander: Plug-and-play inference acceleration for video large language models. (2025). arXiv preprint arXiv:2505.14454. https://arxiv.org/abs/2505.14454
58. H. Gao, Y. Bao, X. Tu, B. Zhong, M. Zhang, Apvr: Hour-level long video understanding with adaptive pivot visual information retrieval. (2025). arXiv preprint arXiv:2506.04953. https://arxiv.org/abs/2506.04953
59. J. Ye, Z. Wang, H. Sun, K. Chandrasegaran, Z. Durante, C. Eyzaguirre, Y. Bisk, J.C. Niebles, E. Adeli, L. Fei-Fei et al., Re-thinking temporal search for long-form video understanding. in *Proceedings of the Computer Vision and Pattern Recognition Conference* (CVPR). (Computer Vision Foundation / IEEE, Nashville, TN, USA 2025). pp. 8579–8591
60. L. Yao, H. Wu, K. Ouyang, Y. Zhang, C. Xiong, B. Chen, X. Sun, J. Li, Generative frame sampler for long video understanding. (2025). arXiv preprint arXiv:2503.09146. https://arxiv.org/abs/2503.09146
61. S. Yu, C. Jin, H. Wang, Z. Chen, S. Jin, Z. Zuo, X. Xu, Z. Sun, B. Zhang, J. Wu et al., Frame-voyager: Learning to query frames for video large language models. (2024). arXiv preprint arXiv:2410.03226. https://arxiv.org/abs/2410.03226
62. W. Guo, Z. Chen, S. Wang, J. He, Y. Xu, J. Ye, Y. Sun, H. Xiong, Logic-in-frames: Dynamic keyframe search via visual semantic-logical verification for long video understanding. (2025). arXiv preprint arXiv:2503.13139. https://arxiv.org/abs/2503.13139
63. B. Fang, W. Wu, Q. Wu, Y. Song, A.B. Chan, Threading keyframe with narratives: Mllms as strong long video comprehenders. (2025). arXiv preprint arXiv:2505.24158. https://arxiv.org/abs/2505.24158
64. L. Gao, Y. Zhong, Y. Zeng, H. Tan, D. Li, Z. Zhao, Linvt: Empower your image-level large language model to understand videos. (2024). arXiv preprint arXiv:2412.05185. https://arxiv.org/abs/2412.05185

65. X. Wang, Y. Zhang, O. Zohar, S. Yeung-Levy, Videoagent: Long-form video understanding with large language model as agent. in *European Conference on Computer Vision* (Springer, Cham 2024). pp. 58–76
66. Z. Yang, D. Chen, X. Yu, M. Shen, C. Gan, Vca: Video curious agent for long video understanding. (2024). arXiv preprint arXiv:2412.10471. https://arxiv.org/abs/2412.10471
67. Y. Fan, X. Ma, R. Wu, Y. Du, J. Li, Z. Gao, Q. Li, Videoagent: A memory-augmented multimodal agent for video understanding. in *European Conference on Computer Vision* (Springer, Cham 2024). pp. 75–92
68. R. Liao, M. Erler, H. Wang, G. Zhai, G. Zhang, Y. Ma, V. Tresp, Videoinsta: Zero-shot long video understanding via informative spatial-temporal reasoning with llms. (2024). arXiv preprint arXiv:2409.20365. https://arxiv.org/abs/2409.20365
69. Z. Wang, B. Chen, Z. Yue, Y. Wang, Y. Qiao, L. Wang, Y. Wang, Videochat-a1: Thinking with long videos by chain-of-shot reasoning. (2025). arXiv preprint arXiv:2506.06097. https://arxiv.org/abs/2506.06097
70. J. Hu, Z. Cheng, C. Si, W. Li, S. Gong, Cos: Chain-of-shot prompting for long video understanding. (2025). arXiv preprint arXiv:2502.06428. https://arxiv.org/abs/2502.06428
71. C. Zhang, Y.B. Lin, Z. Wang, M. Bansal, G. Bertasius, Silvr: A simple language-based video reasoning framework. (2025). arXiv preprint arXiv:2505.24869. https://arxiv.org/abs/2505.24869
72. J. Pereira, V. Lopes, D. Semedo, J. Neves, Self-res: Self-reflection in large vision-language models for long video understanding. (2025). arXiv preprint arXiv:2503.20362. https://arxiv.org/abs/2503.20362
73. Y. Suo, F. Ma, L. Zhu, T. Wang, F. Rao, Y. Yang, From trial to triumph: Advancing long video understanding via visual context sample scaling and self-reward alignment. (2025). arXiv preprint arXiv:2503.20472. https://arxiv.org/abs/2503.20472
74. Y. Wang, L. Zhang, J. Liu, J. Yan, Z. Zhang, J. Zheng, X. Yang, D. Wu, X. Chen, X. Li, Episodic memory representation for long-form video understanding. (2025). arXiv preprint arXiv:2508.09486.
75. Y. Luo, X. Zheng, X. Yang, G. Li, H. Lin, J. Huang, J. Ji, F. Chao, J. Luo, R. Ji, Video-rag: Visually-aligned retrieval-augmented long video comprehension. (2024). arXiv preprint arXiv:2411.13093. https://arxiv.org/abs/2411.13093
76. Z. Xue, J. Zhang, X. Xie, Y. Cai, Y. Liu, X. Li, D. Tao, Omni-adavideorag: Omni-contextual adaptive retrieval-augmented for efficient long video understanding. (2025). arXiv preprint arXiv:2506.13589. https://arxiv.org/abs/2506.13589
77. P. Wang, S. Bai, S. Tan, S. Wang, Z. Fan, J. Bai, K. Chen, X. Liu, J. Wang, W. Ge et al., Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. (2024). arXiv preprint arXiv:2409.12191. https://arxiv.org/abs/2409.12191
78. S. Bai, K. Chen, X. Liu, J. Wang, W. Ge, S. Song, K. Dang, P. Wang, S. Wang, J. Tang et al., Qwen2. 5-vl technical report. (2025). arXiv preprint arXiv:2502.13923. https://arxiv.org/abs/2502.13923
79. B. Li, Y. Zhang, D. Guo, R. Zhang, F. Li, H. Zhang, K. Zhang, P. Zhang, Y. Li, Z. Liu et al., Llava-onevision: Easy visual task transfer. (2024). arXiv preprint arXiv:2408.03326. https://arxiv.org/abs/2408.03326
80. Y. Zhang, Y. Lu, T. Wang, F. Rao, Y. Yang, L. Zhu, Flexselect: Flexible token selection for efficient long video understanding. (2025). arXiv preprint arXiv:2506.00993. https://arxiv.org/abs/2506.00993
81. A. Hurst, A. Lerer, A.P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford et al., Gpt-4o system card. (2024). arXiv preprint arXiv:2410.21276. https://arxiv.org/abs/2410.21276
82. H. Lu, Y. Huo, G. Yang, Z. Lu, W. Zhan, M. Tomizuka, M. Ding, Uniadapter: Unified parameter-efficient transfer learning for cross-modal modeling. in *The Twelfth International Conference on Learning Representations* (ICLR). (Vienna, Austria, OpenReview.net 2024).
83. S. Palaskar, O. Rudovic, S. Dharur, F. Pesce, G. Krishna, A. Sivaraman, J. Berkowitz, A.H. Abdelaziz, S. Adya, A. Tewfik, Multimodal large language models with fusion low rank adaptation for device directed speech detection. (2024). arXiv preprint arXiv:2406.09617. https://arxiv.org/abs/2406.09617
84. S. Ghafouri, M. Fayyaz, X. Li, D. John, B. Ji, D. Nikolopoulos, H. Vandierendonck, Polymorph: Energy-efficient multi-label classification for video streams on embedded devices. (2025). arXiv preprint arXiv:2507.14959. https://arxiv.org/abs/2507.14959
85. E.H. Kim, M.J. Lim, J.H. Shin, Mmer-lmf: multi-modal emotion recognition in lightweight modality fusion. Electronics **14**(11), 2139 (2025)
86. C. Zhu, B. Xiao, L. Shi, S. Xu, X. Zheng, Customize segment anything model for multi-modal semantic segmentation with mixture of lora experts. (2024). arXiv preprint arXiv:2412.04220. https://arxiv.org/abs/2412.04220
87. N. Vouitsis, Z. Liu, S.K. Gorti, V. Villecroze, J.C. Cresswell, G. Yu, G. Loaiza-Ganem, M. Volkovs, Data-efficient multimodal fusion on a single gpu. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR). (Computer Vision Foundation / IEEE, Seattle, WA, USA 2024). pp. 27239–27251
88. Y. Zheng, Y. Chen, B. Qian, X. Shi, Y. Shu, J. Chen, A review on edge large language models: design, execution, and applications. ACM Comput. Surv. **57**(8), 1–35 (2025)
89. X. Wang, W. Jia, Optimizing edge ai: a comprehensive survey on data, model, and system strategies. (2025). arXiv preprint arXiv:2501.03265. https://arxiv.org/abs/2501.03265
90. E. Saurez, H. Gupta, H. Roger, S. Bhowmik, U. Ramachandran, K. Rothermel, Utility-aware load shedding for real-time video analytics at the edge. (2023). arXiv preprint arXiv:2307.02409. https://arxiv.org/abs/2307.02409
91. J. Shao, X. Zhang, J. Zhang, Task-oriented communication for edge video analytics. IEEE Trans. Wirel. Commun. **23**(5), 4141–4154 (2023)
92. H. Zhao, G.P. Ji, R. Yan, H. Xiong, Z. Li, Videoexpert: Augmented llm for temporal-sensitive video understanding. (2025). arXiv preprint arXiv:2504.07519. https://arxiv.org/abs/2504.07519
93. K. Hu, P. Wu, F. Pu, W. Xiao, Y. Zhang, X. Yue, B. Li, Z. Liu, Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. (2025). arXiv preprint arXiv:2501.13826. https://arxiv.org/abs/2501.13826
94. W. Dai, A. Luo, Z. Durante, D. Dash, A. Milstein, K. Schulman, E. Adeli, L. Fei-Fei, Towards fine-grained video question answering. (2025). arXiv preprint arXiv:2503.06820. https://arxiv.org/abs/2503.06820
95. D. Tao, M. Song, X. Li, J. Shen, J. Sun, X. Wu, C. Faloutsos, S.J. Maybank, Bayesian tensor approach for 3-d face modeling. IEEE Trans. Circuits Syst. Video Technol. **18**(10), 1397–1410 (2008)
96. F. Gao, D. Tao, X. Gao, X. Li, Learning to rank for blind image quality assessment. IEEE Trans. Neural Netw. Learn. Syst. **26**(10), 2275–2290 (2015)
97. Z. Chen, S. Dong, K. Yi, Y. Li, M. Ding, A. Torralba, J.B. Tenenbaum, C. Gan, Compositional physical reasoning of objects and events from videos. IEEE Trans. Pattern Anal. Mach. Intell. **47**(9), 7689-7703 (2025) https://doi.org/10.1109/TPAMI.2025.3574322
98. J. Niu, Y. Li, Z. Miao, C. Ge, Y. Zhou, Q. He, X. Dong, H. Duan, S. Ding, R. Qian et al., Ovo-bench: How far is your video-llms from real-world online video understanding? in *Proceedings of the Computer Vision and Pattern Recognition Conference* (CVPR). (Computer Vision Foundation / IEEE, Nashville, TN, USA 2025). pp. 18902–18913
99. H. Zhang, Y. Wang, Y. Tang, Y. Liu, J. Feng, J. Dai, X. Jin, Flash-vstream: Memory-based real-time understanding for long video streams. (2024). arXiv preprint arXiv:2406.08085. https://arxiv.org/abs/2406.08085
100. M.M. Islam, N. Ho, X. Yang, T. Nagarajan, L. Torresani, G. Bertasius, Video recap: Recursive captioning of hour-long videos. in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (CVPR). (Computer Vision Foundation / IEEE, Seattle, WA, USA 2024). pp. 18198–18208
101. J. Yang, S. Liu, H. Guo, Y. Dong, X. Zhang, S. Zhang, P. Wang, Z. Zhou, B. Xie, Z. Wang et al., Egolife: Towards egocentric life assistant. in *Proceedings of the Computer Vision and Pattern Recognition Conference* (CVPR). (Computer Vision Foundation / IEEE, Nashville, TN, USA 2025). pp. 28885–28900
102. S. Huang, H. Zhang, X. Li, Enhance vision-language alignment with noise. in *Proceedings of the AAAI Conference on Artificial Intelligence*, (AAAI). (Philadelphia, PA, USA, AAAI Press 2025). **39**, pp. 17449–17457
103. J. Yan, S. Xu, X. Chen, D. Zhang, J. Tang, G. Wu, J. Liu, Rasr: Retrieval-augmented super resolution for practical reference-based image restoration. (2025). arXiv preprint arXiv:2508.09449. https://arxiv.org/abs/2508.09449
104. Y. Wang, L. Wang, C. Zhang, Y. Zhang, Z. Zhang, A. Ma, C. Fan, T.L. Lam, J. Hu, Learning robust stereo matching in the wild with selective

mixture-of-experts. (2025). arXiv preprint arXiv:2507.04631. https://arxiv.org/abs/2507.04631
105. J. Liu, Z. Ji, Y. Yu, J. Cao, Y. Pang, J. Han, X. Li, Parameter-efficient fine-tuning for continual learning: A neural tangent kernel perspective. (2024). arXiv preprint arXiv:2407.17120. https://arxiv.org/abs/2407.17120
106. J. Liu, Z. Ji, X. Chen, How can diffusion models evolve into continual generators? (2025). arXiv preprint arXiv:2505.11936. https://arxiv.org/abs/2505.11936
107. J. Liu, Z. Ji, Y. Pang, Y. Yu, Ntk-guided few-shot class incremental learning. IEEE Trans. Image Process. **33**, 6029-6044 (2024) https://doi.org/10.1109/TIP.2024.3478854

Publisher's Note

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.