Video-Oasis: Rethinking Evaluation of Video Understanding

Geuntaek Lim[1*], Sungjune Park[1], Jaeyun Lee[1], Inwoong Lee[2], Taeoh Kim[2], Dongyoon Wee[2], Minho Shim[2†], and Yukyung Choi[1†]

[1] Sejong University, South Korea
[2] NAVER Cloud, South Korea
gtlim,ykchoi@rcv.sejong.ac.kr

Abstract. The inherent complexity of video understanding makes it difficult to determine whether Video-LLM benchmark performance stems from visual perception, linguistic reasoning, or knowledge priors. While many benchmarks have emerged to assess high-level reasoning, shared criteria for evaluating video understanding remain largely overlooked. Instead of introducing yet another benchmark, we take a step back to re-examine the criteria for evaluating video understanding. In this work, we introduce Video-Oasis, a sustainable diagnostic suite for systematically auditing existing video understanding benchmarks. This audit reveals that 55% of existing benchmark samples are solvable without visual input or temporal context. After filtering these shortcuts, the remaining video-native challenges expose a substantial capability gap: state-of-the-art models perform only marginally above random guessing. Building on these findings, we use the distilled challenges as a testbed to investigate which algorithmic design choices contribute to robust video understanding. We hope our work provides a practical foundation for constructing rigorous video benchmarks and evaluating future Video-LLMs. Code is available at https://github.com/sejong-rcv/Video-Oasis.

Keywords: Video Understanding · Video-LLM · Benchmark Auditing

1 Introduction

The rise of multi-modal language models has steered video understanding from specific tasks toward the integration of perception and reasoning. While earlier benchmarks focused on narrow domains such as action recognition or temporal localization, video large language models (Video-LLMs) [1, 5, 18, 29, 37] are now required to handle both fine-grained dynamics [9, 16, 21] and long-form reasoning [12, 39, 50]. This expansion, however, makes it difficult to determine whether reported benchmark performance stems from visual perception, linguistic reasoning, or knowledge priors. As benchmarks continue to proliferate across diverse tasks, a unified set of video-centric criteria becomes increasingly necessary for both benchmark creators and users.

* Work done while doing an internship at NAVER Cloud.
† Co-corresponding authors.

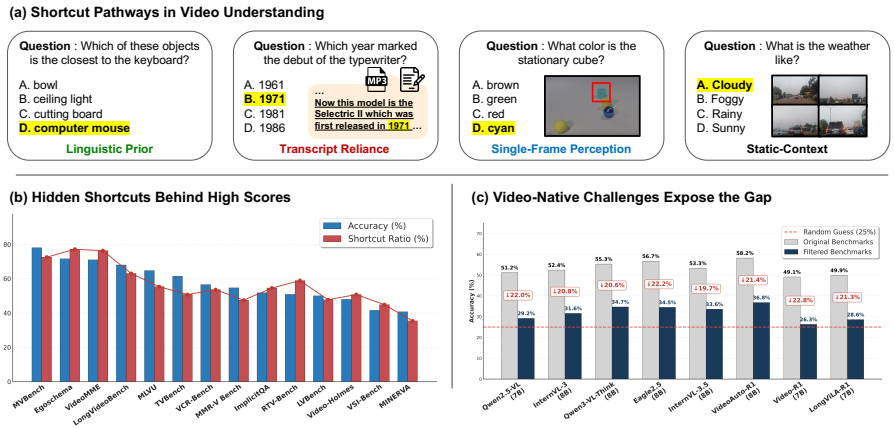

Fig. 1: (a) Existing video-QA benchmarks include shortcut-solvable instances that can be answered without spatio-temporal video understanding. (b) Benchmarks with higher shortcut ratios tend to report higher accuracy. (c) State-of-the-art Video-LLMs consistently exhibit a substantial drop when facing video-native challenges, revealing the inherent difficulty of robust spatio-temporal understanding.

In this work, rather than introducing yet another benchmark, we take a step back to re-examine the essential criteria required for evaluating video understanding. As illustrated in Figure 1(a), existing benchmarks often include samples that do not require visual evidence or temporal context, but can instead be solved through linguistic priors, audio cues, or static visual evidence. To address this issue, we introduce **Video-Oasis**, a diagnostic suite for auditing whether existing video understanding benchmarks truly require visual and temporal dependencies. Video-Oasis consists of three key components: (i) visual-dependency tests that remove or replace visual evidence to identify samples solvable without grounded perception; (ii) temporal-dependency tests that perturb or remove temporal order to identify samples solvable without temporal reasoning; and (iii) ambiguity verification that uses human-in-the-loop inspection to identify annotation issues arising from the complexity of video data. Together, these components provide a systematic way to filter shortcut-solvable samples and distill video-native challenges such as temporal continuity, causal interaction, and multi-event narratives.

With this diagnostic suite, we conduct a large-scale audit of 14 diverse benchmarks [7,9,12,16,21,22,27,33,35,39,42,43,50,52], covering tasks from perception to reasoning and video durations from seconds to hours. Specifically, we decouple visual and temporal cues through a series of diagnostic tests and define the ***shortcut ratio*** as the proportion of samples within a benchmark that can be solved without visual or temporal dependency. As shown in Figure 1(b), reported benchmark accuracy is strongly correlated with shortcut prevalence, suggesting that high scores can be partially inflated by samples that do not require robust video understanding.

Table 1: Comparison with prior studies auditing benchmarks. **Evaluation Scope** denotes the number of benchmarks analyzed. **Diagnostic Coverage** indicates whether visual, temporal, and ambiguity-related diagnostic axes are considered, where • denotes multiple tests, ◦ denotes a single test, and – denotes not considered. **Cross-Model Consensus** indicates whether conclusions are derived using an ensemble of models, and **Manual Verification** denotes whether human verification is performed.

Previous Work	Evaluation Scope	Diagnostic Coverage			Cross-Model Consensus	Manual Verification
		Vis.	Tem.	Amb.		
EgoTempo [26]	2	–	◦	–	–	–
Cambrian-S [44]	9	•	◦	–	–	–
Apollo [53]	6	◦	◦	–	–	✓
Video-Oasis (Ours)	**14**	•	•	•	✓	✓

The Video-Oasis audit reveals two key findings: (i) **High shortcut prevalence:** 55% of existing benchmark samples are solvable without visual input or temporal context; (ii) **Limited performance on video-native challenges:** after filtering these shortcuts, state-of-the-art models [1, 6, 19, 36] perform only marginally above random chance, as shown in Figure 1(c). These findings indicate that current benchmarks can overestimate models' video understanding capabilities, while the remaining video-native challenges expose substantial limitations in current models.

The broad audit scope and multi-axis diagnostic design of Video-Oasis distinguish it from prior benchmark-auditing studies. As shown in Table 1, prior studies have identified important issues, including temporal shortcuts [26], perception-oriented task design [44], and benchmark redundancy [53]. Video-Oasis extends these efforts by jointly examining visual dependency, temporal dependency, and ambiguity across 14 benchmarks. It further incorporates cross-model consensus and human verification to improve the reliability of the diagnostic process.

The remainder of this paper is organized as follows. We first introduce the Video-Oasis diagnostic suite and audit existing benchmarks in Section 3. We then analyze the distilled video-native challenges and evaluate a broad range of video understanding models in Section 4. Finally, we use these challenges as a testbed to investigate algorithmic design choices for robust video understanding in Section 5. In this work, we make the following major contributions:

- **Revisiting Video Understanding.** We revisit the fundamental criteria that video understanding benchmarks should satisfy and identify a critical gap in current benchmark design.
- **Holistic Diagnostic Framework.** We propose Video-Oasis, a rigorous and sustainable diagnostic framework that jointly examines visual dependency, temporal dependency, and ambiguity to distill video-native challenges.
- **Practical Guidelines.** Through comprehensive analyses enabled by Video-Oasis, we derive practical guidelines for benchmark construction and the algorithmic design of future video understanding models.

2 Related Work

2.1 Large Language Models for Video Understanding

The evolution of Video-LLMs [1, 5, 18, 29, 37] has centered on aligning high-dimensional visual features with the semantic space of LLMs. Early research employed temporal pooling [18] or attention mechanisms [37] to compress video frames into token sequences within the model's context window. More recently, the field has shifted toward scaling context windows [5, 29, 31] to process longer video sequences without aggressive temporal compression. Despite these advancements, Video-LLMs often struggle with complex temporal narratives due to their fixed, feed-forward processing paradigm. To address this, dynamic agentic frameworks [3, 4, 30, 38, 41, 45, 46, 49] have emerged as a complementary paradigm, utilizing structured task decomposition and iterative reasoning loops. While the synergy between Video-LLMs and agentic frameworks has led to rapid gains on existing benchmarks [12, 21, 39, 50], it remains unclear whether these gains stem from visual perception, linguistic reasoning, or knowledge priors.

2.2 Challenges in Video Benchmark Construction

As Video-LLMs [5, 19] and agentic methods [3, 10, 46] rapidly improve performance on existing benchmarks [12, 21, 39, 50], evaluation benchmarks must be constructed with greater rigor to properly attribute these gains. However, the inherent complexity of video understanding, combined with the large volume of data, often makes the construction of evaluation datasets challenging. Consequently, dataset construction pipelines frequently rely on automatic strategies, such as generating questions from selected keyframes [14, 15] or using LLMs to produce questions based on video transcripts [16, 39, 43]. In this process, it becomes unclear whether the resulting benchmarks truly evaluate video-specific properties that distinguish video from other modalities, such as temporal continuity, causal interaction, and multi-event narratives. This gap calls for a rigorous re-examination of whether current pipelines preserve essential video-specific dependencies and whether reported gains reflect spatio-temporal reasoning.

2.3 Toward Robust Video Understanding Benchmarks

As benchmark proliferation enables comprehensive evaluation, it also introduces unintended issues. Apollo [53] highlights the redundancy across existing benchmarks, while Cambrian-S [44] observes that many tasks remain overly concentrated on perception-oriented evaluation. In line with these works [26, 44, 53], we take a step back to re-examine the current landscape of video understanding. Building upon these efforts, we introduce Video-Oasis, a diagnostic suite that distills existing datasets to isolate core spatio-temporal challenges while suppressing unintended shortcuts through visual–temporal decoupling tests. By integrating cross-model consensus and human-in-the-loop verification, Video-Oasis extends prior analyses by providing a sustainable framework for auditing modern video understanding benchmarks.

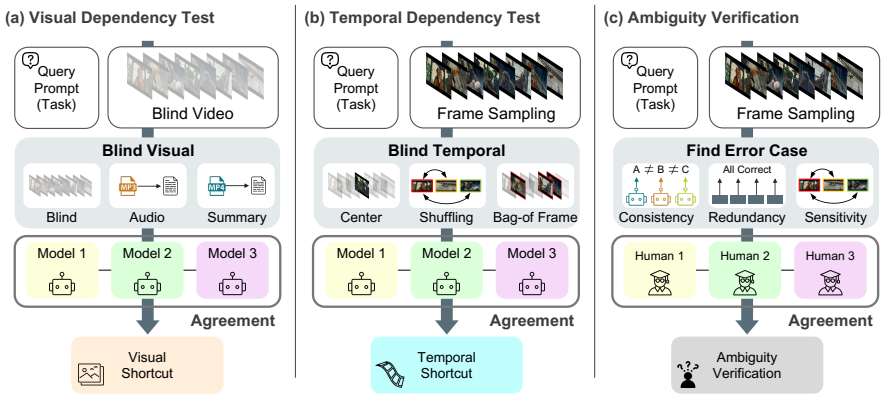

Fig. 2: Overview of the Video-Oasis diagnostic suite, which assesses (a) whether visual information is required, (b) whether temporal context is necessary, and (c) whether the task contains ambiguity in video data.

3 Video-Oasis: Diagnostic Suite for Video Understanding

Establishing reliable protocols for measuring spatio-temporal reasoning remains a critical yet underexplored challenge in video understanding. To address this, we introduce Video-Oasis, a diagnostic suite for verifying whether video benchmarks satisfy shared criteria for video understanding. First, we introduce our test design, which systematically examines the essential dependencies required for video understanding (Section 3.1). Next, we audit existing benchmarks using our diagnostic protocols (Section 3.2). Finally, we examine the diagnostic coverage of Video-Oasis and the validity of shortcut identification (Section 3.3). Experimental settings are detailed in Sec. C of the supplementary material.

3.1 Design of the Diagnostic Suite

Criteria 1: Is Visual Evidence Required? To test visual dependency, we replace the original video with inputs that remove or abstract away raw visual evidence. As illustrated in Figure 2(a), we use three diagnostic tests: (i) **Blind**, which provides only the question and answer options to identify cases solvable through linguistic bias or world knowledge without any visual input; (ii) **Audio**, where the video's audio track is transcribed into text and given to the model in place of the video; and (iii) **Summary**, where the raw video is replaced by a concatenated sequence of captions [40] extracted at fixed intervals.

The Audio and Summary tests are intentionally simple and are not designed to optimize text-only video reasoning. Although recent methods have explored treating video reasoning as long-document comprehension through iterative retrieval or memory updating [20, 38], we use these textual inputs for a different purpose. They serve as diagnostic probes: if a sample can be answered from an audio transcript or caption sequence, it may not require grounded visual perception from the raw video.

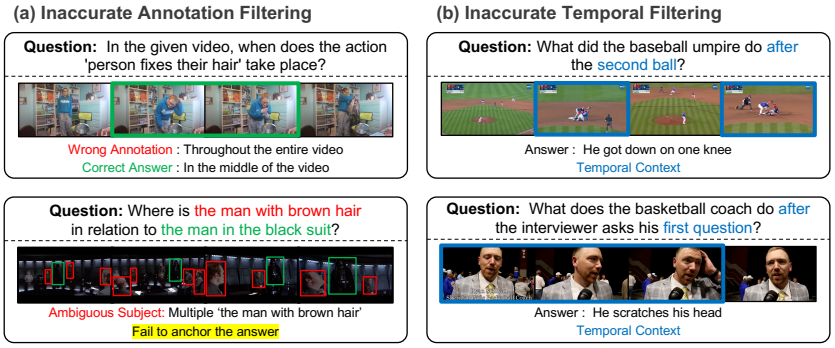

Fig. 3: (a) Inaccurate annotations identified by the redundancy and consistency tests. (b) Questions incorrectly filtered by the frame shuffling test but manually restored.

Criteria 2: Is Temporal Context Required? Temporal dependency is central to video understanding because many questions require ordering events, tracking state changes, or reasoning about causal interactions. As illustrated in Figure 2(b), we verify this dependency through three diagnostic tests: (i) **Center-Frame**, which provides only the middle frame to detect if the task functions merely as spatial recognition without temporal depth; (ii) **Frame Shuffling**, where we randomly permute the frame order to disrupt temporal causality and measure the model's sensitivity to chronological sequences; and (iii) **Bag-of-Frames (BoF)**, which employs a frozen CLIP-based encoder [28, 32, 48] that does not model temporal order to perform top-k frame matching against the query. If this non-temporal, similarity-based approach succeeds, the task does not necessitate temporal reasoning but rather simple visual pattern recognition.

Criteria 3: Is the Annotation Reliable? Since video data are long and information-dense, video-QA annotations can be ambiguous, involving imprecise temporal grounding, incomplete evidence, or non-unique answers. While many existing dataset construction pipelines often overlook this uncertainty, we apply three checks that flag samples for manual inspection, as illustrated in Figure 2(c): (i) **Consistency**, identifying cases where models [1, 5, 11, 19, 36] fail to reach a consensus, which suggests inherent ambiguity or non-unique answers; (ii) **Redundancy**, investigating cases solvable via any arbitrary video segment, revealing flawed question designs or global biases that fail to anchor the answer to a specific temporal segment; and (iii) **Sensitivity**, where we manually verify cases in which models succeed despite frame shuffling to account for potential uncertainty in the sequential understanding of Video-LLMs.

Figure 3 provides representative examples of the manual verification process. Figure 3(a) shows annotation issues identified by the consistency and redundancy checks, including incorrect temporal labels and ambiguous subjects that prevent the answer from being reliably anchored to the video. Figure 3(b) shows cases initially filtered by the frame-shuffling test, but manually restored because the questions still require temporal ordering, such as reasoning about what happens after a specific event.

3.2 Benchmarking the Benchmarks

Diagnostic Test Results. We conduct comprehensive diagnostic tests across existing benchmarks, leveraging a diverse set of Video-LLMs [1, 2, 5, 19] and VLMs [28, 32, 48]. The results in Table 2 are aggregated over all 14 benchmarks. Notably, models achieve accuracies ranging from 30% to 50% even when raw visual evidence or temporal order is removed or disrupted, compared to the random-chance baseline of 25.6%. These results suggest that many benchmark samples do not strictly enforce the intended visual or temporal dependencies. Additional robustness analysis and benchmark-wise results are provided in Secs. A and B of the supplementary material, respectively.

Video-LLM	Acc.
Eagle2.5	35.6
Qwen2.5-VL	33.5
Qwen3-VL	36.2

(a) Blind Test

Video-LLM	Acc.
Eagle2.5	47.6
Qwen2.5-VL	46.8
Qwen3-VL	45.9

(b) Audio

Video-LLM	Acc.
Eagle2.5	44.0
Qwen2.5-VL	42.6
Qwen3-VL	45.0

(c) Summary

Video-LLM	Acc.
Eagle2.5	42.2
Qwen3-VL	40.2
VideoAuto-R1	43.0

(d) Center-Frame

Video-LLM	Acc.
Eagle2.5	52.2
Qwen3-VL	50.7
VideoAuto-R1	52.4

(e) Frame Shuffling

VLM	Acc.
CLIP	31.4
Long-CLIP	32.8
EVA-CLIP	32.7

(f) Bag-of-Frames

Table 2: Aggregate diagnostic-test results over 14 benchmarks. The metric is accuracy.

Auditing Existing Benchmarks. To conduct a granular analysis based on task characteristics, we manually group the 14 benchmarks into spatial, temporal, reasoning, and general categories. We define the consensus threshold (c) as the number of diagnostic models listed in Table 2 that must answer a sample correctly for it to be counted as a shortcut. Under this framework, we aggregate shortcut instances across all tests and compare their ratios under different consensus thresholds. Table 3 reveals two key findings: (i) shortcut-solvable samples appear consistently across all task groups, and (ii) under a relaxed consensus threshold ($c \geq 1$), an average of 92.7% of samples exhibit shortcut-solvable behavior under at least one visual or temporal diagnostic test. These results indicate that many existing benchmarks do not sufficiently enforce the spatio-temporal dependencies required for video understanding.

Table 3: Ratio of shortcuts (%) across comprehensive benchmark groups.

Consensus Threshold	Spatial [21, 33, 43]	Temporal [9, 27, 42]	Reasoning [7, 22, 52]	General [12, 16, 35, 39, 50]
$c \geq 1$	95.6	95.7	85.8	94.0
$c \geq 2$	86.1	85.2	69.2	83.9
$c = 3$	58.8	54.4	44.6	63.0

3.3 Empirical Insights into the Diagnostic Framework

Diagnostic Test Distribution. Video-Oasis includes ambiguity tests to refine unreliable or uncertain samples before constructing the final filtered set. As summarized in Table 4, the consistency and redundancy checks address unreliable annotations, while the sensitivity check corrects potential false positives from frame shuffling tests. We then analyze how shortcuts are distributed across the individual diagnostic tests of Video-Oasis. Under the strict consensus condition ($c = 3$), Table 5 reports both total and unique shortcut counts for each diagnostic test. Since the tests are not perfectly orthogonal, the unique counts measure the distinct contribution of each test beyond overlaps with others. Table 5 reveals two findings: (i) *Summary*, *Center-Frame*, and *Frame Shuffling* account for the majority of unique shortcuts (65%), forming a practical protocol for future benchmark construction; and (ii) *Blind*, *Audio*, and *Bag-of-Frames* still contribute a non-negligible 35%, demonstrating that Video-Oasis provides complementary diagnostic coverage.

Table 4: Statistics of manual refinement.

Ambiguity Test	# of Samples	
	Total	Refined
Consistency	666	213
Redundancy	477	197
Sensitivity	1,758	804

Table 5: Statistics of diagnostic tests.

Diagnostic Test	# of Samples	
	Total	Unique
Blind	2,751	362
Audio	1,685	301
Summary	6,703	1,308
Center-Frame	5,882	847
Frame Shuffling	8,280	1,758
Bag-of-Frames	4,309	1,394

Validity of Shortcut Identification. Because shortcut identification relies on restricted settings that remove or weaken visual or temporal evidence, success in these settings may not verify shortcut behavior. To assess the validity of shortcut identification, we define the correlation rate as the fraction of shortcut-identified samples that are correctly solved under standard evaluation with uniformly sampled frames. We use different models [6,10,36] from those used for shortcut identification to avoid circular validation. As Table 6 shows, the identified samples exhibit a high correlation rate, averaging 76% across tests and models. This indicates that the identified shortcut cases are already handled well by current models, suggesting that future evaluation should move beyond them and focus on the remaining problems that continue to challenge current models.

Table 6: Correlation rate (%) of shortcut-identified samples under standard evaluation.

Models	**Blind**	**Audio**	**Summary**	**Center Frame**	**Bag-of Frames**	**Frame Shuffling**
InternVL-3.5 (8B) [36]	77.0	74.0	79.2	79.4	67.2	84.1
LongViLA-R1 (7B) [6]	78.8	72.8	78.1	77.7	65.7	81.3
STAR [10]	74.0	79.8	76.4	74.5	68.0	76.5

4 Distilling the Challenges of Video Understanding

We next examine the challenges that remain after filtering shortcut-solvable samples and evaluate how state-of-the-art models perform on these video-native challenges. Specifically, we first identify the types of video-native challenges distilled by Video-Oasis (Section 4.1), and then evaluate state-of-the-art models under this distilled setting to reveal the remaining gap in strict spatio-temporal understanding (Section 4.2).

4.1 Understanding the Video-Native Challenges

After applying Video-Oasis, the remaining samples are more likely to require strong visual and temporal dependencies. We further analyze these samples to understand what types of video-native challenges they represent. To this end, rather than manually inspecting each sample, we design an efficient pipeline that leverages existing annotations.

Specifically, we first aggregate the source-benchmark metadata associated with the surviving QA pairs, including their original task categories and sub-task labels. Using this metadata, we prompt Gemini-2.5-Pro [13] to derive candidate challenge clusters by abstracting the remaining tasks under the Video-Oasis criteria. We then consolidate these initial clusters into five unified categories that capture the dominant capabilities required by the Video-Oasis-filtered samples.

- **Fine-Grained Perception:** requires grounding fine-grained recognition in a spatio-temporal context, where visual identification depends on how details evolve across space and time.
- **Spatial World Understanding:** requires synthesizing fragmented, multi-view evidence across frames to infer 3D context, including relative positions, geometry, and trajectories.
- **Temporal Dynamics & Tracking:** requires monitoring changes over time, such as object tracking, action sequencing, and state transitions, demanding temporal ordering to prevent reliance on unordered frame matching.
- **Causality & Logical Reasoning:** requires deducing latent cause-and-effect relationships, physical laws, and unobserved intentions, going beyond pixel-level observations to probe the implicit logic of the video.
- **Global Narrative:** requires integrating events across the full timeline to infer long-term semantics or overarching plots while filtering out irrelevant contexts over extended video horizons.

Next, we assign each surviving QA pair to a single primary challenge category. For this categorization step, each annotator model receives the question, answer options, and the definitions of the five categories. To improve annotation consistency, we employ an ensemble of five proprietary LLMs [23–25]. A category label is accepted when at least three models agree; otherwise, the sample is manually inspected and labeled. Only 122 samples fail to reach consensus among the five LLMs and require manual inspection. Further details on category annotation prompts and robustness to annotation-model choices are provided in Sec. D.3 of the supplementary material.

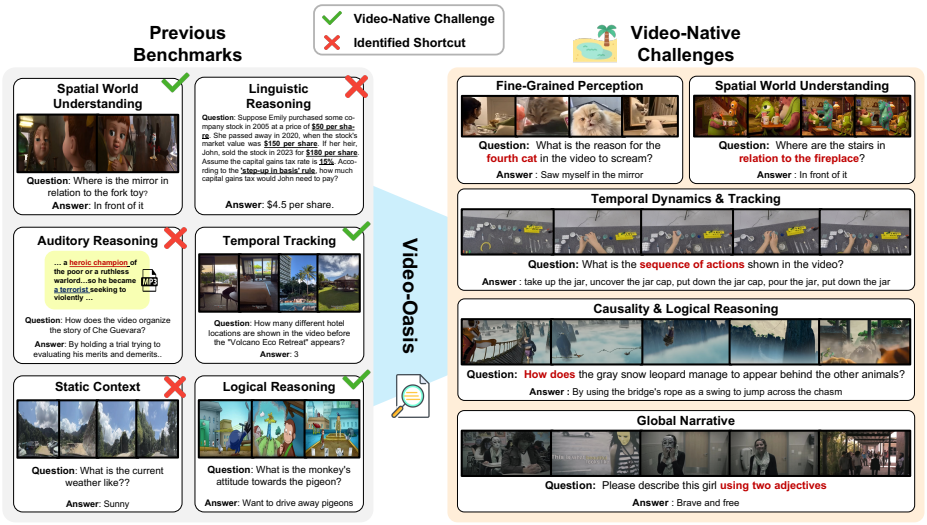

Fig. 4: Video-Oasis filters shortcut-driven samples from existing benchmarks and distills video-native challenges that require fine-grained perception, spatial understanding, temporal tracking, causal reasoning, and global narrative understanding.

Our goal is not to introduce novel tasks, but to refine the fundamental capabilities of video understanding. While these categories may resemble traditional taxonomies, they are derived from a bottom-up, data-driven process. By filtering out shortcut-solvable samples, these video-native challenges are derived from the remaining tasks rather than imposed as predefined benchmark categories.

Overview. From the 14 curated benchmarks covering diverse aspects of video understanding, 11,033 QA pairs remain from the original 24,416, associated with 4,938 unique videos. By reducing the evaluation volume by 55%, Video-Oasis enables more efficient evaluation while preserving essential spatio-temporal challenges. The diagnostic tests used to construct this set are reproducible and extensible, enabling seamless incorporation of new benchmarks and emerging challenges. Additional statistics on the distilled video-native challenges are provided in Sec. D.1 of the supplementary material.

Qualitative Examples. Video-Oasis establishes a rigorous evaluation setting that specifically targets strict spatio-temporal dependencies. In Figure 4(left), tasks solvable through simple frame-level perception or linguistic/auditory reasoning without visual input fail to satisfy video-specific criteria. Moving beyond simple frame-level recognition, Video-Oasis highlights challenges such as spatial and semantic matching across disparate views, chronological reasoning, and the preservation of temporal continuity, as illustrated in Figure 4(right). These scenarios represent inherently video-native challenges that remain difficult to solve without access to visual and temporal evidence, highlighting the core directions that future video understanding evaluation should emphasize. More qualitative examples and category-level explanations are provided in Sec. D.2 of the supplementary material.

Beyond Difficulty-Based Filtering. We further examine whether the distilled video-native challenges are merely difficult questions for current models. To this end, we compare the Video-Oasis-distilled set with a difficulty-based baseline of 6,609 questions that Eagle2.5 [5], Qwen3-VL [1], and VideoAuto-R1 [19] all answer incorrectly. The two sets show only 44.6% overlap. This indicates that Video-Oasis does not simply collect hard questions, but identifies samples through predefined visual, temporal, and ambiguity diagnostics that better reflect video-native dependencies.

4.2 Comprehensive Evaluation

Experimental Settings. We conduct an extensive evaluation across a diverse spectrum of models, encompassing open-source Video-LLMs [1,5,11,19], leading proprietary models [13,23], and agentic methods [10,38]. To ensure a fair comparison across all models, the input visual context is restricted to a maximum of 128 frames, uniformly sampled at a rate of 1 fps. For agentic frameworks [10,38], we adhere to the default configurations in their original implementations, except that the reasoning model is replaced with GPT-5-mini [25]. To ensure experimental reliability, we provide a detailed validation comparing official benchmark scores with our reproduced results in Sec. E of the supplementary material.

Table 7: Benchmarking state-of-the-art models under video-native challenges.

Models	Fine. Percep.	Spatial World	Temporal Dynamics	Causal Reason.	Global Narrat.	Overall
Proprietary LLMs						
GPT-4o [23]	25.6	33.2	26.3	27.3	26.5	27.5
Gemini-2.5-Pro [13]	**40.2**	**49.8**	**50.9**	**45.4**	**43.0**	**46.7**
Open-Source Video-LLMs						
Qwen2.5-VL (7B) [2]	23.3	28.7	32.3	28.6	21.2	29.2
Qwen3-VL$_{\text{inst.}}$(8B) [1]	27.0	42.4	36.5	28.0	21.5	33.8
Qwen3-VL$_{\text{think.}}$(8B) [1]	29.0	41.6	37.7	27.7	23.2	34.6
Eagle2.5 (8B) [5]	26.9	31.0	39.7	33.2	22.7	34.5
InternVL-3 (8B) [51]	27.0	31.3	34.1	30.6	24.5	31.6
InternVL-3.5 (8B) [36]	29.5	41.9	35.1	29.8	23.3	33.6
Video-R1 (7B) [11]	24.0	24.0	29.1	27.3	18.4	26.3
LongViLA-R1 (7B) [6]	28.4	25.4	31.5	27.9	20.6	28.6
VideoAuto-R1 (8B) [19]	27.5	44.3	39.5	31.1	28.9	36.8
Agentic Methods						
VideoTree (GPT-5$_{\text{mini}}$) [38]	28.6	34.0	32.3	24.6	20.7	30.1
STAR (GPT-5$_{\text{mini}}$) [10]	<u>31.6</u>	<u>44.4</u>	<u>42.2</u>	<u>34.0</u>	<u>32.9</u>	<u>39.5</u>

Experimental Results. Comprehensive quantitative results for all methods under the distilled evaluation setting are shown in Table 7. We report category-wise accuracy for each model and use aggregate accuracy over all samples as the overall score. From these results, we draw several conclusions:

1. **Performance nearing random-chance levels:** A key observation is that current Video-LLMs [2, 6, 11] exhibit performance close to the chance level (25.6%), with the exception of a few top-tier models [10, 13], confirming that current models struggle with rigorous spatio-temporal reasoning.
2. **The current frontier of video reasoning:** Gemini-2.5-Pro [13] achieves the highest performance across all skills; however, its moderate performance also indicates that these tasks pose a non-trivial challenge even for the most advanced proprietary models.
3. **Bottleneck in holistic understanding:** The consistently low performance in the *Global Narrative* dimension reveals that long-term multi-scene understanding remains a primary bottleneck for current architectures.
4. **Impact of agentic designs:** Agentic design choices, such as VideoTree [38] and STAR [10], significantly influence overall efficacy even when using the same reasoning model [25], underscoring the importance of reasoning-step orchestration.

5 Exploring Algorithmic Designs

The distilled video-native challenges expose a substantial gap in current Video-LLMs. We next use this setting to examine which algorithmic designs can help close this gap. Because these challenges emphasize visual and temporal dependencies, they provide a focused testbed for analyzing model improvements.

5.1 The Role of Temporal Grounding

Temporal grounding associates a query with the video segments needed to answer it, making it especially critical for video-native challenges with strict spatio-temporal dependencies. To investigate its practical impact, we conduct an ablation study using AKS [34] to retrieve the 16 most relevant frames as a representative temporal grounding method. As shown in Table 8, integrating this grounding pipeline yields consistent improvements across models. However, the magnitude of these gains remains modest, with improvements of 1.4% for Eagle2.5 [5] and 2.3% for Qwen3-VL-Instruct [1], despite their relatively low baseline performance. This raises a natural question: are the limited gains caused by weak reasoning, or by imperfect temporal grounding?

Table 8: Ablation study of temporal grounding in Video-LLMs. The metric is accuracy.

Method	Temporal Grounding	Fine. Perception	Spatial World	Temporal Dynamics	Causal Logical	Global Narrative	Overall
Eagle2.5 [5]	-	25.0	29.4	34.9	30.5	25.7	31.5
	✓	28.4	31.9	35.7	30.4	27.5	**32.9**
Qwen3-VL-Instruct [1]	-	22.9	37.1	28.8	24.8	16.9	27.8
	✓	25.2	37.9	32.3	23.3	19.9	**30.1**

To isolate the impact of grounding from uncertainty, we conducted an oracle experiment on 2,945 QA pairs from ImplicitQA [33] and KFS-Bench [17].

The set contains 1,060 Video-Oasis-distilled samples and 1,885 shortcut samples, each paired with ground-truth temporal regions. Table 9 reveals a clear contrast: while oracle grounding leads to a substantial performance gain on Video-Oasis-distilled samples (35.0% → 50.8%), the gain in shortcuts is notably smaller (78.0% → 80.8%), confirming that samples identified as shortcuts often allow models to bypass precise temporal grounding. This finding yields a crucial insight: **precise grounding becomes increasingly important in environments where strong spatio-temporal dependencies are required.**

Table 9: Upper bound performance (%) with oracle temporal grounding.

Method	Fine. Perception	Spatial World	Temporal Dynamics	Causal Logical	Global Narrative	Video-Oasis (overall)	Shortcut (overall)
Eagle2.5	37.2	22.9	40.5	38.5	16.0	35.0	78.0
Eagle2.5 (w. Oracle)	50.4	27.5	61.3	48.1	48.0	**50.8**	**80.8**

5.2 Adaptive Reasoning: When to Think and When to Perceive

Video-Oasis emphasizes strict spatio-temporal dependencies that often require multi-hop understanding. This makes reasoning-depth control an important design axis: models must decide not only how to perceive the video, but also when deeper reasoning is necessary. To this end, we employ Qwen3-VL (8B) [1] as the base model and compare: (i) the instruction-following mode, (ii) the thinking mode, and (iii) adaptive thinking via VideoAuto-R1 [19]. As detailed in Table 10, adaptive thinking (Qwen3-VL$_{\text{AutoR1}}$) outperforms an always-on thinking strategy (Qwen3-VL$_{\text{think.}}$), suggesting that reasoning depth should be adjusted to the question rather than fixed in advance.

Table 10: Ablation study of reasoning depth modulation. The metric is accuracy.

Method	Fine. Perception	Spatial World	Temporal Dynamics	Causal Logical	Global Narrative	Overall
Qwen3-VL$_{\text{inst.}}$ [1]	27.0	42.4	36.5	28.0	21.5	33.8
Qwen3-VL$_{\text{think.}}$ [1]	29.0	41.6	37.7	27.7	23.2	34.6
Qwen3-VL$_{\text{AutoR1}}$ [19]	27.5	44.3	39.5	31.1	28.9	36.8
Qwen3-VL$_{\text{voting}}$	38.4	**57.7**	49.4	37.8	30.1	46.2
Gemini-2.5-Pro [13]	**40.2**	49.8	**50.9**	**45.4**	**43.0**	**46.7**

We further ask how much performance could improve if the model chose the better mode for each question. To explore this, we introduce an oracle ensemble baseline (Qwen3-VL$_{\text{voting}}$), which considers a response correct if either the instruction-following (Qwen3-VL$_{\text{inst.}}$) or the thinking mode (Qwen3-VL$_{\text{think.}}$) successfully solves the task. By simulating an optimal selection between thinking and non-thinking states, the voting baseline reaches 46.2, nearly closing the gap with the frontier-level Gemini-2.5-Pro (46.7). This finding yields a crucial insight: **the strategic optimization of when to think can be as impactful as the raw scale of the model's architecture.**

5.3 Training Paradigms

In this section, we review different training paradigms and conduct systematic experiments to compare their effectiveness for video understanding. Specifically, our empirical analysis is twofold: (i) a direct performance comparison between supervised fine-tuning (SFT) and reinforcement learning with verifiable rewards (RLVR), and (ii) a targeted investigation into RLVR reward designs to determine the critical components for complex video understanding. Our empirical evaluation, leveraging Qwen2.5-VL [2] as the unified base model in Table 11, yields the following key insights:

1. **Effectiveness of long-context SFT:** Eagle2.5 [5] demonstrates a clear margin of improvement over the baseline Qwen2.5-VL [2], improving overall accuracy from 29.2% to 34.5% without relying on RLVR. This indicates that advanced long-context optimization can serve as a key factor in enhancing general spatio-temporal reasoning.
2. **RLVR Reward Designs:** The comparative results among RLVR-based models and the SFT baseline reveal a non-linear performance trend: Video-R1 [11] (26.3%) < Qwen2.5-VL [2] (29.2%) < VideoAuto-R1$_{\text{Qwen2.5}}$ [19] (32.7%). These divergent outcomes indicate that reward formulation is critical to the success of RL-based video training. Despite its importance, the systematic investigation of reward structures for video reasoning remains largely underexplored, representing a potential area for future research.
3. **Complementary Strengths of SFT and RLVR:** Our results suggest that SFT and RLVR offer complementary strengths rather than a single superior path. While well-optimized SFT, as represented by Eagle2.5 [5], improves overall accuracy, RLVR with grounding rewards, as represented by VideoAuto-R1 [19], provides larger gains on specific reasoning challenges such as *Global Narrative* (21.2%→28.6%). Recent studies [8, 47] similarly explore the complementary roles of SFT and RLVR, but remain largely confined to linguistic reasoning or static image domains, motivating their extension to video understanding.

Table 11: Comparison of SFT and RLVR training paradigms for video understanding. All models share the same base LLM, Qwen2.5-VL [2]. The metric is accuracy.

Models	Rewards		Fine. Perception	Spatial World	Temporal Dynamics	Causal Logical	Global Narrative	Overall
	QA	Grounding						
Qwen2.5-VL [2]	-	-	23.3	28.7	32.3	28.6	21.2	29.2
Eagle2.5 [5]	-	-	**26.9**	**31.0**	**39.7**	<u>33.2</u>	<u>22.7</u>	**34.5**
Video-R1 [11]	✓	-	24.0	24.0	29.1	27.3	18.4	26.3
VideoAuto-R1$_{\text{Qwen2.5}}$ [19]	✓	✓	<u>25.4</u>	<u>29.7</u>	<u>35.9</u>	**33.7**	**28.6**	<u>32.7</u>

In summary, Video-Oasis not only diagnoses limitations in evaluation protocols but also reveals key insights into the algorithmic components of video understanding through comprehensive ablation studies. These findings provide practical guidance for designing stronger video understanding methods.

6 Conclusion

In this work, we establish Video-Oasis, a rigorous diagnostic suite for robust video understanding. Through this diagnostic lens, we analyze vulnerabilities in existing benchmarks with respect to visual and temporal dependency and re-examine the current landscape of video understanding. Beyond diagnosis, our algorithmic exploration provides several insights, highlighting temporal grounding and adaptive reasoning as primary drivers of spatio-temporal reasoning. We further identify the balance between SFT and RLVR as an open question for future Video-LLM training. Our work is fully reproducible, and we open-source the entire pipeline to enable large-scale auditing of existing datasets and provide an extensible evaluation protocol for the community. We hope Video-Oasis serves as a foundation for developing more rigorous benchmarks and drives the next generation of models toward robust video understanding.

Discussion. Diagnostic tests such as caption-based or shuffled-frame evaluation may preserve partial temporal cues, leading to false positives or false negatives. To reduce this risk, Video-Oasis combines complementary diagnostic axes with cross-model consensus and human verification. We further view Video-Oasis as a reproducible and configurable audit suite that can be adapted to different evaluation goals by adjusting its tests.

Acknowledgements

This work was supported by the NAVER Cloud Corporation and partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MSIT) (RS-2025-00553785, 40%), the IITP under the virtual convergence support program to nurture the best talents grant funded by the Korea government (MSIT) (IITP-2026-RS-2023-00254529, 40%), and the IITP under the Leading Generative AI Human Resources Development grant funded by the Korea government (MSIT) (IITP-2026-RS-2026-25544647, 20%).

References

1. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
2. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
3. Chen, B., Wang, Z., Yue, Z., Yan, K., Yu, C., Huang, Y., Liu, Z., Wen, Y., Chen, X., Liu, Y., et al.: Videochat-m1: Collaborative policy planning for video understanding via multi-agent reinforcement learning. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2026)
4. Chen, B., Yue, Z., Chen, S., Wang, Z., Liu, Y., Li, P., Wang, Y.: Lvagent: Long video understanding by multi-round dynamical collaboration of mllm agents. In: Proceedings of the International Conference on Computer Vision (2025)

5. Chen, G., Li, Z., Wang, S., Jiang, J., Liu, Y., Lu, L., Huang, D.A., Byeon, W., Le, M., Ehrlich, M., Lu, T., Wang, L., Catanzaro, B., Kautz, J., Tao, A., Yu, Z., Liu, G.: Eagle 2.5: Boosting long-context post-training for frontier vision-language models. In: Proceedings of the Neural Information Processing Systems (2025)
6. Chen, Y., Huang, W., Shi, B., Hu, Q., Ye, H., Zhu, L., Liu, Z., Molchanov, P., Kautz, J., Qi, X., Liu, S., Yin, H., Lu, Y., Han, S.: Scaling RL to long videos. In: Proceedings of the Neural Information Processing Systems (2025)
7. Cheng, J., Ge, Y., Wang, T., Ge, Y., Liao, J., Shan, Y.: Video-holmes: Can mllm think like holmes for complex video reasoning? arXiv preprint arXiv:2505.21374 (2025)
8. Chu, T., Zhai, Y., Yang, J., Tong, S., Xie, S., Schuurmans, D., Le, Q.V., Levine, S., Ma, Y.: SFT memorizes, RL generalizes: A comparative study of foundation model post-training. In: Proceedings of the International Conference on Machine Learning (2025)
9. Cores, D., Dorkenwald, M., Mucientes, M., Snoek, C.G., Asano, Y.M.: Lost in time: A new temporal benchmark for videollms. In: Proceedings of the British Machine Vision Conference (2025)
10. Fan, S., Cui, J., Guo, M.H., Yang, S.: Tool-augmented spatiotemporal reasoning for streamlining video question answering task. In: Proceedings of the Neural Information Processing Systems (2025)
11. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in MLLMs. In: Proceedings of the Neural Information Processing Systems (2025)
12. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)
13. Google DeepMind, Comanici, G., Bieber, E., et al.: Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. arXiv preprint arXiv:2507.06261 (2025)
14. Han, S., Huang, W., Shi, H., Zhuo, L., Su, X., Zhang, S., Zhou, X., Qi, X., Liao, Y., Liu, S.: Videoespresso: A large-scale chain-of-thought dataset for fine-grained video reasoning via core frame selection. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)
15. Li, K., He, Y., Wang, Y., Li, Y., Wang, W., Luo, P., Wang, Y., Wang, L., Qiao, Y.: Videochat: Chat-centric video understanding. arXiv preprint arXiv:2305.06355 (2023)
16. Li, K., Wang, Y., He, Y., Li, Y., Wang, Y., Liu, Y., Wang, Z., Xu, J., Chen, G., Luo, P., et al.: Mvbench: A comprehensive multi-modal video understanding benchmark. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2024)
17. Li, Z., Ishida, K., Yamazaki, S., Ji, X., Liu, J.: Kfs-bench: Comprehensive evaluation of key frame sampling in long video understanding. In: Proceedings of the Winter Conference on Applications of Computer Vision (2026)
18. Lin, B., Ye, Y., Zhu, B., Cui, J., Ning, M., Jin, P., Yuan, L.: Video-llava: Learning united visual representation by alignment before projection. In: Proceedings of the Empirical Methods in Natural Language Processing (2024)
19. Liu, S., Zhuge, M., Zhao, C., Chen, J., Wu, L., Liu, Z., Zhu, C., Cai, Z., Zhou, C., Liu, H., et al.: Videoauto-r1: Video auto reasoning via thinking once, answering twice. arXiv preprint arXiv:2601.05175 (2026)

20. Ma, Z., Gou, C., Shi, H., Sun, B., Li, S., Rezatofighi, H., Cai, J.: Drvideo: Document retrieval based long video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)
21. Mangalam, K., Akshulakov, R., Malik, J.: Egoschema: A diagnostic benchmark for very long-form video language understanding. In: Proceedings of the Neural Information Processing Systems (2023)
22. Nagrani, A., Menon, S., Iscen, A., Buch, S., Mehran, R., Jha, N., Hauth, A., Zhu, Y., Vondrick, C., Sirotenko, M., et al.: Minerva: Evaluating complex video reasoning. In: Proceedings of the International Conference on Computer Vision (2025)
23. OpenAI: GPT-4o system card. arXiv preprint arXiv:2410.21276 (2024)
24. OpenAI: OpenAI o3 and o4-mini system card. https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf (2025)
25. OpenAI: GPT-5 system card. arXiv preprint arXiv:2601.03267 (2026)
26. Plizzari, C., Tonioni, A., Xian, Y., Kulshrestha, A., Tombari, F.: Omnia de egotempo: Benchmarking temporal understanding of multi-modal llms in egocentric videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)
27. Qi, Y., Zhao, Y., Zeng, Y., Bao, X., Huang, W., Chen, L., Chen, Z., Zhao, J., Qi, Z., Zhao, F.: Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. arXiv preprint arXiv:2504.07956 (2025)
28. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: Proceedings of the International Conference on Machine Learning (2021)
29. Shen, X., Xiong, Y., Zhao, C., Wu, L., Chen, J., Zhu, C., Liu, Z., Xiao, F., Varadarajan, B., Bordes, F., Liu, Z., Xu, H., Kim, H.J., Soran, B., Krishnamoorthi, R., Elhoseiny, M., Chandra, V.: LongVU: Spatiotemporal adaptive compression for long video-language understanding. In: Proceedings of the International Conference on Machine Learning (2025)
30. Shen, X., Zhang, W., Chen, J., Elhoseiny, M.: Vgent: Graph-based retrieval-reasoning-augmented generation for long video understanding. In: Proceedings of the Neural Information Processing Systems (2025)
31. Shu, Y., Liu, Z., Zhang, P., Qin, M., Zhou, J., Liang, Z., Huang, T., Zhao, B.: Video-xl: Extra-long vision language model for hour-scale video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)
32. Sun, Q., Fang, Y., Wu, L., Wang, X., Cao, Y.: Eva-clip: Improved training techniques for clip at scale. arXiv preprint arXiv:2303.15389 (2023)
33. Swetha, S., Gupta, R., Kulkarni, P.P., Shatwell, D.G., Santiago, J.A.C., Siddiqui, N., Fioresi, J., Shah, M.: Implicitqa: Going beyond frames towards implicit video reasoning. arXiv preprint arXiv:2506.21742 (2025)
34. Tang, X., Qiu, J., Xie, L., Tian, Y., Jiao, J., Ye, Q.: Adaptive keyframe sampling for long video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)
35. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., Qi, J., Ding, M., Gu, X., Huang, S., Xu, B., et al.: Lvbench: An extreme long video understanding benchmark. In: Proceedings of the International Conference on Computer Vision (2025)
36. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., et al.: Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)

37. Wang, Y., Li, K., Li, X., Yu, J., He, Y., Chen, G., Pei, B., Zheng, R., Wang, Z., Shi, Y., et al.: Internvideo2: Scaling foundation models for multimodal video understanding. In: Proceedings of the European Conference on Computer Vision (2024)
38. Wang, Z., Yu, S., Stengel-Eskin, E., Yoon, J., Cheng, F., Bertasius, G., Bansal, M.: Videotree: Adaptive tree-based video representation for llm reasoning on long videos. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 3272–3283 (2025)
39. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. In: Proceedings of the Neural Information Processing Systems (2024)
40. Xu, Y., Li, X., Yang, Y., Meng, D., Huang, R., Wang, L.: Carebench: A fine-grained benchmark for video captioning and retrieval. In: Proceedings of the International Conference on Learning Representations (2026)
41. Xue, Z., Zhang, J., Xie, X., yuxuan cai, Liu, Y., Li, X., Tao, D.: AdavideoRAG: Omni-contextual adaptive retrieval-augmented efficient long video understanding. In: Proceedings of the Neural Information Processing Systems (2025)
42. Xun, S., Tao, S., Li, J., Shi, Y., Lin, Z., Zhu, Z., Yan, Y., Li, H., Zhang, L., Wang, S., Liu, Y., Zhang, H., Ma, Y., Hu, X.: RTV-bench: Benchmarking MLLM continuous perception, understanding and reasoning through real-time video. In: Proceedings of the Neural Information Processing Systems (2025)
43. Yang, J., Yang, S., Gupta, A.W., Han, R., Fei-Fei, L., Xie, S.: Thinking in space: How multimodal large language models see, remember, and recall spaces. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)
44. Yang, S., Yang, J., Huang, P., II, E.L.B., Yang, Z., Yu, Y., Tong, S., Zheng, Z., Xu, Y., Wang, M., Fergus, R., LeCun, Y., Fei-Fei, L., Xie, S.: Towards spatial supersensing in video. In: Proceedings of the International Conference on Learning Representations (2026)
45. Yeo, W., Kim, K., Yoon, J., Hwang, S.J.: Worldmm: Dynamic multimodal memory agent for long video reasoning. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2026)
46. Yin, X., Peng, X., Li, X., Xiong, Z., Lu, Y.: Hierarchical long video understanding with audiovisual entity cohesion and agentic search. arXiv preprint arXiv:2601.13719 (2026)
47. Yue, Y., Chen, Z., Lu, R., Zhao, A., Wang, Z., Yue, Y., Song, S., Huang, G.: Does reinforcement learning really incentivize reasoning capacity in LLMs beyond the base model? In: Proceedings of the Neural Information Processing Systems (2025)
48. Zhang, B., Zhang, P., Dong, X., Zang, Y., Wang, J.: Long-clip: Unlocking the long-text capability of clip. In: Proceedings of the European Conference on Computer Vision (2024)
49. Zhang, X., Jia, Z., Guo, Z., Li, J., Li, B., Li, H., Lu, Y.: Deep video discovery: Agentic search with tool use for long-form video understanding. In: Proceedings of the Neural Information Processing Systems (2025)
50. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., Xiao, S., Qin, M., Yang, X., Xiong, Y., Zhang, B., et al.: Mlvu: Benchmarking multi-task long video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)
51. Zhu, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)
52. Zhu, K., Jin, Z., Yuan, H., Li, J., Tu, S., Cao, P., Chen, Y., Liu, K., Zhao, J.: MMR-v: What's left unsaid? a benchmark for multimodal deep reasoning in videos. In: Proceedings of the International Conference on Learning Representations (2026)

53. Zohar, O., Wang, X., Dubois, Y., Mehta, N., Xiao, T., Hansen-Estruch, P., Yu, L., Wang, X., Juefei-Xu, F., Zhang, N., et al.: Apollo: An exploration of video understanding in large multimodal models. In: Proceedings of the Computer Vision and Pattern Recognition Conference (2025)