V-HORIZON: A DIAGNOSTIC BENCHMARK FOR ULTRA-LONG VIDEO UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Despite rapid progress in Video Large Language Models (VideoLLMs), strong benchmark performance does not necessarily reflect genuine visual understanding, especially for videos spanning tens of hours. We introduce **V-Horizon**, a diagnostic benchmark for ultra-long video understanding with 660 six-option multiple-choice questions from 34 complete TV series ranging from 4 to over 50 hours. Rather than relying on a single aggregate score, V-Horizon decomposes performance into four complementary capabilities: **Temporal Tracking** (T), **Narrative Understanding** (N), **Visual Perception** (V), and **Visual Localization** (W), separating text-inferable reasoning from tasks that require explicit *visual grounding*. We evaluate 15 representative models and systems across holistic and agentic paradigms, together with keyframe-selection and visual-token-compression baselines. Our results reveal a substantial gap between text-supported and visually grounded performance: models perform well when linguistic cues suffice, but degrade sharply when they must retrieve sparse visual evidence over long temporal contexts. Oracle-evidence experiments show that visual localization accuracy improves by up to 39 percentage points when the relevant evidence is provided, indicating that **long-range evidence retrieval** is a primary bottleneck in ultra-long video understanding. Motivated by this finding, we propose **Agentic Skills**, which combines hybrid retrieval with local VLM-based analysis and achieves 70.3% accuracy, outperforming the strongest holistic baseline by 15 percentage points. Overall, V-Horizon provides a diagnostic testbed for distinguishing apparent video understanding supported by textual shortcuts from genuine long-range visual grounding.

1 INTRODUCTION

Video Large Language Models (VideoLLMs) have achieved strong performance on a range of established benchmarks (Li et al., 2024; Fu et al., 2025; Wu et al., 2024; Chen et al., 2024a). Yet, it remains unclear to what extent such performance reflects genuine visual understanding. Models can often answer video questions by exploiting subtitles, prior knowledge, or linguistic cues without grounding their predictions in the visual content. This limitation becomes particularly pronounced for ultra-long videos spanning tens of hours, where relevant visual evidence is often sparse and transient: a model must first locate the right moments within a vast temporal context before it can reason over them. Existing long-video benchmarks provide only a partial view of this challenge. Most focus on videos lasting from tens of minutes to a few hours (Fu et al., 2025; Wang et al., 2024a; Zhou et al., 2024), leaving substantially longer temporal horizons underexplored. More importantly, aggregate accuracy does not reveal whether a correct prediction is grounded in visual evidence or can be inferred largely from textual cues. This limits our ability to diagnose what current models actually understand—and where they fail—when reasoning over ultra-long videos.

We introduce **V-Horizon**, a diagnostic benchmark for ultra-long video understanding. V-Horizon comprises 660 calibrated six-option multiple-choice questions drawn from 34 complete TV series spanning 4 to over 50 hours. Rather than treating ultra-long video understanding as a single capability, we decompose it along four complementary axes: **Temporal Tracking (T)**, which measures the ability to track entity states, timelines, and event order; **Narrative Understanding (N)**, which captures plots, causal relations, and story structure; **Visual Perception (V)**, which evaluates fine-grained visual recognition; and **Visual Localization (W)**, which tests whether models can locate relevant

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

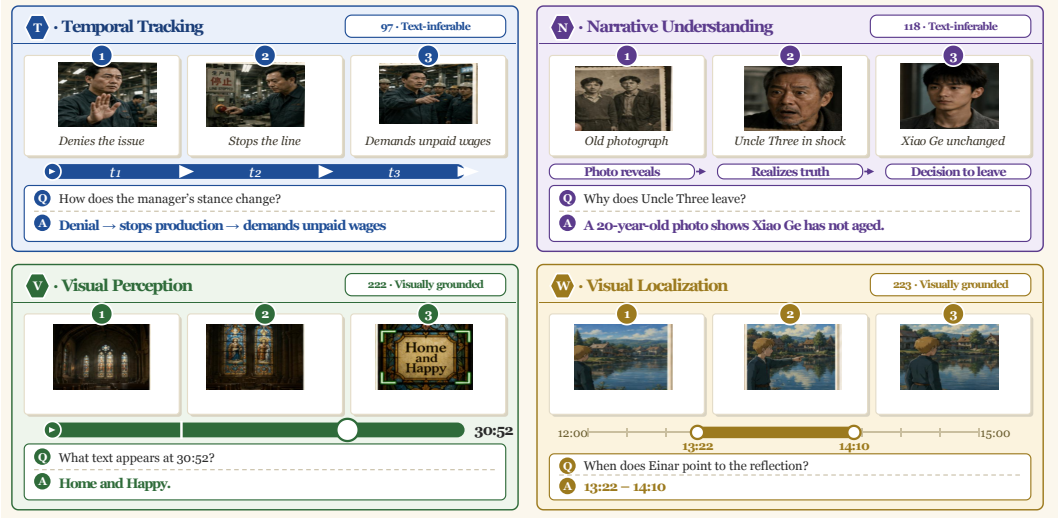


Figure 1: Overview of V-Horizon. Four tasks evaluate long-horizon video understanding across text-inferable and visually grounded settings.

visual events over long temporal contexts. Crucially, this decomposition separates capabilities that are largely *text-inferable* from those that require explicit *visual grounding*, making V-Horizon diagnostic: beyond measuring answer accuracy, it reveals whether models can retrieve and ground the visual evidence required to support their predictions.

We evaluate 15 representative models across holistic and agentic paradigms. Our results reveal a pronounced gap between text-supported reasoning and visually grounded understanding. Holistic models perform substantially better on **Temporal Tracking (T)** and **Narrative Understanding (N)** than on **Visual Perception (V)** and **Visual Localization (W)**. For example, Gemini 3.1 Pro achieves 82.5% accuracy on T and 76.3% on N, but only 41.0% on V and 46.6% on W. This disparity suggests that strong aggregate performance on ultra-long videos can be driven substantially by textual cues rather than robust visual grounding. Further analysis identifies **long-range visual evidence retrieval** as a primary bottleneck. When oracle evidence frames are provided, W accuracy improves by up to 39 percentage points, indicating that many failures occur before fine-grained visual interpretation: models often fail to locate the relevant evidence within the vast temporal context. Once the correct evidence is retrieved, local visual reasoning becomes substantially more reliable. Motivated by this finding, we introduce **Agentic Skills**, a framework in which an LLM orchestrator dynamically invokes specialized tools for multimodal retrieval, temporal navigation, and fine-grained visual inspection. The framework progressively narrows the temporal search space and delegates candidate regions to a lightweight local VLM for detailed analysis. Using Qwen3-VL-8B as the visual inspector, Agentic Skills achieves 70.3% overall accuracy, outperforming the strongest holistic baseline by 15.0 percentage points. Overall, our results suggest that the central challenge of ultra-long video understanding is not simply processing more visual input, but **retrieving sparse, question-relevant visual evidence from long temporal contexts and reasoning over it effectively**. V-Horizon provides a diagnostic benchmark for studying this challenge and for disentangling text-supported reasoning from genuine long-range visual grounding. Our contributions are as follows:

1. We introduce **V-Horizon**, a diagnostic benchmark for ultra-long video understanding, covering complete videos from 4 to over 50 hours and explicitly distinguishing text-inferable reasoning from visually grounded understanding.
2. We reveal a substantial gap between text-supported and visually grounded performance in current VideoLLMs, and identify **long-range visual evidence retrieval** as a primary bottleneck in ultra-long video understanding.
3. We propose **Agentic Skills**, a retrieval-centric framework that combines multimodal retrieval with localized visual analysis, substantially improving performance on video understanding.

Table 1: Comparison with long-form video benchmarks. V-Horizon extends beyond 50 hours with cross-episode reasoning and text–visual grounding.

Benchmark	Max. Duration	# Choices	# Task Types	Cross-Episode	Text–Visual Diagnosis
Video-MME (Fu et al., 2024)	1 h	4	6	×	×
LongVideoBench (Wu et al., 2024)	1 h	4	17	×	×
MLVU (Zhou et al., 2024)	2 h	4	9	×	×
LVBench (Wang et al., 2024a)	2 h	4	6	×	×
CG-Bench (Chen et al., 2025)	1 h	4	3	×	×
HourVideo (Chandrasegaran et al., 2024)	2 h	5	4	×	×
V-Horizon (Ours)	**50+ h**	**6**	**4**	✓	✓

2 Related Work

Video Understanding Benchmarks. Early benchmarks focus on short video clips (Li et al., 2024; Fang et al., 2024), while recent datasets extend to minutes or hours, including Video-MME (Fu et al., 2025), LongVideoBench (Wu et al., 2024), MLVU (Zhou et al., 2024), EgoSchema (Mangalam et al., 2023), LVBench (Wang et al., 2024a), and CG-Bench (Chen et al., 2025). However, video understanding over tens of hours remains largely unexplored, and aggregate accuracy often obscures whether models rely on textual cues or grounded visual evidence. **V-Horizon** extends evaluation to complete series of up to 50+ hours and explicitly distinguishes text-inferable reasoning from visually grounded understanding.

Video Large Multimodal Models. Recent VideoLLMs have substantially improved video perception and reasoning. Proprietary models such as GPT-4 (Achiam et al., 2023) and Gemini (Team et al., 2024) benefit from large context windows, while open-source models including Qwen-VL (Bai et al., 2025), VideoLLaMA3 (Zhang et al., 2025), VideoChat-Flash (Li et al., 2026), and Video-R1 (Feng et al., 2026) continue to advance long-context video modeling. Yet, large context alone does not guarantee effective retrieval and use of visual evidence over long temporal horizons.

Agentic and Tool-Augmented Video Understanding. Agentic methods address long videos through iterative search and tool use. VideoAgent (Wang et al., 2024b) performs LLM-guided navigation, VideoTree (Wang et al., 2025b) builds hierarchical video representations, while ViperGPT (Surís et al., 2023) and ProViQ (Choudhury et al., 2023) decompose visual reasoning into executable steps. These approaches reduce the need for uniform video processing, but still depend on reliably locating sparse, task-relevant evidence. This motivates retrieval-centric strategies that first narrow the temporal search space and then perform focused visual reasoning.

3 The V-Horizon Benchmark

3.1 Design Principles

V-Horizon is built around three core design principles that distinguish it from prior benchmarks (Zhou et al., 2024; Wu et al., 2024; Chandrasegaran et al., 2024). First, it adopts a **diagnostic capability decomposition** rather than collapsing heterogeneous abilities into a single aggregate score. Specifically, **Temporal Tracking (T)** and **Narrative Understanding (N)** are largely *text-inferable*, as they can often be resolved from subtitles and linguistic context, whereas **Visual Perception (V)** and **Visual Localization (W)** are *visually grounded* and require explicit visual evidence. This separation forms the central diagnostic mechanism of V-Horizon. Second, V-Horizon operates at an **ultra-horizon scale**: each instance consists of a complete TV series spanning 4 to over 50 hours, requiring models to locate sparse evidence across long temporal contexts and reason across episodes. Third, V-Horizon is designed for **high discrimination**. It adopts a six-option multiple-choice format, yielding a random-chance accuracy of 16.7%, together with model-in-the-loop difficulty calibration that filters out questions correctly answered by at least 15 baseline models. This removes overly easy or shortcut-prone examples and places greater emphasis on challenging, visually grounded questions where linguistic cues alone are insufficient.

Table 2: V-Horizon capability taxonomy across text-inferable and visually grounded axes.

Axis	Capability	Grounding	Count	Representative Example
T	Temporal Tracking	Text-inferable	97	"How does the factory manager's stance toward his company change throughout Episode 4?"
N	Narrative Understanding	Text-inferable	118	"Why does Uncle Three panic and abruptly leave after seeing a 20-year-old photo?"
V	Visual Perception	Visually grounded	222	"At 30:52 in Episode 16, what words are visible on the stained-glass windows?"
W	Visual Localization	Visually grounded	223	"At what time does Einar tell Thorfinn to look at his reflection in the water?"
	Total		**660**	

3.2 Data Collection and Annotation

Video corpus. We curate a corpus of 34 complete TV series across four genres: documentaries (8), dramas (10), anime (9), and reality shows (7). The series exhibit substantial variation in duration, ranging from 4 to over 50 hours, with a median length of approximately 12 hours. To support consistent downstream processing and retrieval, we apply three standardized preprocessing pipelines to all videos: (1) Whisper-based ASR (Radford et al., 2023) for speech transcription, (2) dense visual captioning at 10-second intervals using a local VLM, and (3) multimodal embedding and indexing, using BGE-M3 (Chen et al., 2024b) for textual representations and LanguageBind (Zhu et al., 2024) for video clips. Importantly, the ASR transcripts and dense visual captions are used solely as intermediate representations for candidate discovery and retrieval, rather than as benchmark ground-truth annotations.

Question generation pipeline. Our synthesis pipeline consists of four stages. (1) *Event-grounded generation*: a language model generates candidate questions from temporally localized events identified using subtitle transcripts and dense visual captions. These representations are used only for candidate discovery, while the final question, answer, and supporting evidence are verified against the raw video. (2) *Distractor synthesis*: for each question, five plausible but incorrect options are generated using temporal, visual, and narrative confusion strategies. (3) *Blind-test filtering*: candidate questions are evaluated under controlled input conditions to assess modality dependence. Questions answerable from the question and options alone are removed, while subtitle-only performance is used to distinguish *text-inferable* questions from those requiring visual evidence. (4) *Difficulty calibration*: we further remove overly easy questions that are correctly answered by at least 15 evaluated baseline models.

After filtering and calibration, the final benchmark contains 660 curated questions across four complementary capability axes, with an intentional emphasis on visually grounded questions. Full details of the synthesis pipeline and modality diagnostics are provided in Appendix C.

3.3 Capability Taxonomy and Benchmark Statistics

V-Horizon comprises 660 curated questions from 34 complete TV series, organized along four capability axes as summarized in Table 2. By question count, the benchmark spans drama (59%), anime (22%), documentary (10%), and reality shows (9%). To validate the distinction between *text-inferable* and *visually grounded* capabilities, we evaluate a blind text-only setting in which GPT-5.6 receives subtitle transcripts without access to visual frames. As detailed in Appendix D, this setting achieves 60.2% on **Temporal Tracking (T)** and 62.5% on **Narrative Understanding (N)**, but only 21.3% on **Visual Perception (V)** and 18.9% on **Visual Localization (W)**. This large gap empirically confirms that V and W depend strongly on visual evidence, whereas T and N remain substantially more accessible from textual cues alone.

4 Agentic Skills: Tool-Augmented Reasoning

To investigate whether long-range evidence retrieval is a primary bottleneck in ultra-long video understanding, we introduce **Agentic Skills**, a tool-augmented reasoning framework. The framework

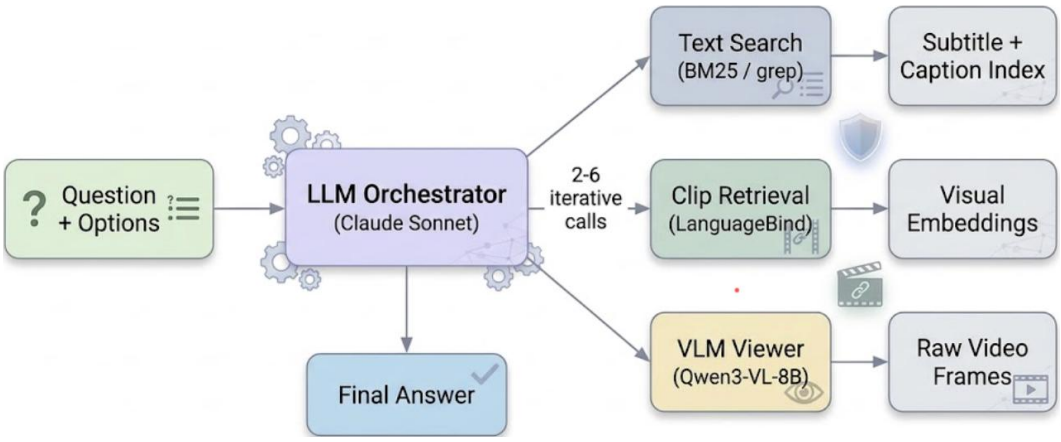


Figure 2: Agentic multimodal QA pipeline with autonomous skill invocation.

equips an LLM orchestrator (Claude Sonnet (Bai et al., 2022)) with three complementary tools for evidence retrieval and verification:

Text Search. We perform lexical retrieval using BM25 and grep over timestamped subtitle transcripts (ASR) and dense visual captions generated at 10-second intervals, returning ranked passages with explicit temporal boundaries.

Visual Clip Retrieval. We perform cross-modal semantic retrieval over pre-computed LanguageBind (Zhu et al., 2024) embeddings to identify video segments relevant to the query.

VLM Viewer. For fine-grained visual verification, the agent extracts a localized clip of up to 120 seconds and queries a lightweight local VLM (Qwen3-VL-8B (Bai et al., 2025)) to inspect the retrieved evidence.

As shown in Figure 2, the orchestrator first interprets the question and formulates a retrieval strategy, then iteratively invokes these tools over two to six turns before producing the final answer, retaining accumulated evidence in context so that later turns can refine the query or resolve ambiguities left by earlier retrieval. Importantly, the VLM Viewer uses only an 8B-parameter model, substantially smaller than frontier holistic models such as Gemini 3.1 Pro and GPT-5.6. This design helps separate the benefits of effective evidence retrieval and localization from those of stronger visual modeling alone, allowing us to directly examine whether better access to relevant evidence drives performance on videos.

5 Experiments

5.1 Experimental Setup

Models and systems. We evaluate 15 representative model and system configurations, organized into three main categories:

- **Agentic and tool-augmented methods:** Agentic Skills (with Claude), Qwen MM Plugins (with Qwen-Max) (Qwen Team, 2026), VideoSeek (Lin et al., 2026), and DVD (Zhang et al., 2026b) (with Qwen3-VL-32B). Unlike conventional one-shot video understanding models, these methods incorporate additional mechanisms such as retrieval, structured memory, tool use, and iterative video navigation to actively gather and refine task-relevant evidence.
- **Closed-source models:** Gemini 3.1 Pro (Team et al., 2023), GPT-5.6, Gemini 3 Flash (Team et al., 2023), Doubao Seed 2.0 (Seed, 2026), and GPT-4o (Hurst et al., 2024).
- **Open-source models:** Qwen3.6-35B (Yang et al., 2025), VideoChat-Flash (Li et al., 2026), Qwen3-VL-8B (Bai et al., 2025), Video-R1-7B (Feng et al., 2026), VideoLLaMA3-7B (Zhang et al., 2025), and InternVL3.5-8B (Wang et al., 2025a).

Strategy	Budget Allocation	Model Input	Coverage–Continuity Trade-off
Fixed-frame	Fixed frame budget over the full series	Timestamped JPEG frames / contact sheets	Broad coverage but no local motion; per-episode coverage decreases with series length.
Fixed-segment	Up to 10 segments over the full series	Ordered low-FPS MP4 segments	Preserves local motion and shot transitions, but leaves temporal gaps between segments.
Token-budgeted	FPS adapted to duration, resolution, and token budget	One low-FPS MP4 per episode	Preserves within-episode continuity, but sampling becomes sparser for longer series.

Table 3: Comparison of global video-input strategies. The approaches differ in how they trade off global temporal coverage and local continuity under a finite visual budget.

In addition to these model and system baselines, we evaluate four training-free input-processing strategies to isolate the benefits of efficient video evidence selection and compression. We apply each strategy to Qwen3-VL-8B and InternVL3.5-8B. At the frame level, the **CLIP-based** baseline (Ye et al., 2023) selects question-relevant frames based on cross-modal similarity, while **Adaptive Keyframe Sampling (AKS)** (Tang et al., 2025) jointly accounts for question–frame relevance and temporal coverage to identify a compact yet informative set of keyframes. At the visual-token level, **CDPruner** (Zhang et al., 2026a) removes redundant tokens while preserving a diverse, question-conditioned subset, whereas **PruMerge** (Shang et al., 2025) combines token pruning and merging to reduce spatial redundancy. Together, these baselines represent two complementary efficiency paradigms: *temporal sparsification*, which reduces the number of frames to be processed, and *spatial/token compression*, which reduces the number of visual tokens within the retained inputs.

Although these configurations differ substantially in how they access and process video evidence, we evaluate each system using the input interface and inference procedure for which it is designed, with the goal of fully utilizing its native input capabilities. For example, some models directly consume video or sampled visual inputs, whereas agentic and tool-augmented systems may retrieve, revisit, or iteratively navigate relevant video content. We therefore do not force all systems into a single input representation; instead, we allow each configuration to operate in its intended setting so that the comparison reflects its strongest practical video-understanding capability.

All configurations are evaluated on the same 660-question V-Horizon test set, with identical questions, six answer options, and output-format instructions. Thus, while the mechanisms for accessing video evidence may differ across systems, the prediction task and evaluation criterion remain fixed. The six-way random baseline is 16.7%. We report exact-match accuracy after parsing each response into one of the six option labels.

Global video construction. We preserve each model's native long-video input interface rather than enforcing a unified preprocessing pipeline. As shown in Table 3, the resulting implementations fall into three broad families.

Fixed-frame methods operate on uniformly sampled, timestamped JPEG frames. For each TV series we first build a canonical pool of 2,048 frames, split across episodes as evenly as possible, and then subsample this pool to each model's input limit. GPT-5.6 receives 256 of these frames composed into 16 contact sheets, each a 4×4 grid whose tiles carry visible episode and timestamp labels; the remaining fixed-frame models receive individual base64 JPEG frames, between 48 and 1,024 per series, interleaved in chronological order with their episode and timestamp. Because the frame budget is fixed at the series level, effective per-episode coverage necessarily thins out as the number of episodes or their total duration increases.

Gemini *fixed-segment* models receive up to 10 pre-generated low-frame-rate MP4 segments that together span the full series. In contrast to isolated frames, these segments preserve short-range motion and shot transitions; if the combined segments overflow the context window, the number of segments is halved and the remaining segments are re-selected evenly across the series.

Table 4: **Main results on V-Horizon** (660 six-way multiple-choice questions). **T/N** denote text-dominant question types, whereas **V/W** require explicit visual grounding. **T+N** and **V+W** are question-count-weighted pooled accuracies. **Gap**= $\mathrm{Acc}_{\mathrm{T+N}} - \mathrm{Acc}_{\mathrm{V+W}}$ measures the signed text–vision disparity in percentage points; proximity to zero indicates balance, not absolute capability. *f*, VT, and *seg.* denote frames, visual tokens, and video segments, respectively.

Model / Method	Setting	Overall↑	Question Type				Pooled Accuracy		
			T↑	N↑	V↑	W↑	T+N↑	V+W↑	Gap
Agentic and Tool-Augmented Methods									
Agentic Skills	–	70.3	87.6	73.7	60.4	70.9	80.0	65.6	14.4
Qwen MM Plugins	–	60.0	73.2	78.0	49.1	55.6	75.8	52.4	23.5
DVD	–	37.3	49.5	50.9	34.7	27.4	50.2	31.0	19.2
VideoSeek	–	34.9	46.4	49.2	32.9	24.2	47.9	28.5	19.4
Closed-Source Video Models									
Gemini 3.1 Pro	10 *seg.*	55.3	82.5	76.3	41.0	46.6	79.1	43.8	35.3
GPT-5.6	256*f*	53.9	75.3	84.8	43.2	39.0	80.5	41.1	39.3
Gemini 3 Flash	10 *seg.*	50.3	65.0	74.6	36.0	45.3	70.2	40.7	29.6
Doubao Seed 2.0	256*f*	37.0	49.5	57.6	31.1	26.5	54.0	28.8	25.2
GPT-4o	48*f*	22.0	23.7	27.1	20.7	19.7	25.6	20.2	5.4
Open-Source Video Models									
Qwen3.6-35B	512*f*	42.1	59.8	65.3	33.3	30.9	62.8	32.1	30.7
	220K VT	48.5	66.0	69.5	46.0	32.3	67.9	39.1	28.8
VideoChat-Flash	512*f*	37.6	55.5	53.4	29.7	31.4	54.3	30.6	23.7
	240K VT	38.9	49.5	58.5	31.1	31.8	54.4	31.5	23.0
Qwen3-VL-8B	512*f*	36.3	36.1	50.8	27.5	37.4	44.2	32.5	11.7
	120K VT	36.5	44.3	50.9	33.3	28.7	47.9	31.0	16.9
Video-R1-7B	512*f*	25.0	34.0	41.5	18.9	18.4	38.1	18.7	19.5
	1024*f*	35.6	49.5	42.4	25.2	36.3	45.6	30.8	14.8
VideoLLaMA3-7B	512*f*	34.2	45.4	48.3	23.9	32.3	47.0	28.1	18.9
	1024*f*	30.3	40.2	43.2	24.3	25.1	41.9	24.7	17.1
InternVL3.5-8B	512*f*	31.5	41.2	50.8	25.7	22.9	46.5	24.3	22.2
	240K VT	28.5	33.0	27.1	32.0	23.8	29.8	27.9	1.9
Frame / Token Selection Baselines									
Qwen3-VL-8B	CDPruner	34.7	39.2	41.5	36.5	27.4	40.5	31.9	8.6
	CLIP	34.1	42.3	44.1	31.1	28.3	43.3	29.7	13.6
	AKS	33.2	40.2	48.3	27.0	28.3	44.7	27.6	17.0
	PruMerge	32.4	36.1	47.5	27.9	27.4	42.3	27.6	14.7
InternVL3.5-8B	CLIP	29.8	36.1	46.6	23.9	24.2	41.9	24.0	17.8
	AKS	27.7	36.1	43.2	24.3	19.3	40.0	21.8	18.2
	CDPruner	27.4	27.8	36.4	26.1	23.8	32.6	24.9	7.6
	PruMerge	25.5	28.9	33.1	23.9	21.5	31.2	22.7	8.5

vLLM-based models use *token-budgeted video*. A series-level sampling rate is derived from the model's visual-token budget, the target resolution, and the total series duration: each frame is charged a fixed number of visual tokens determined by its spatial patch count (two consecutive frames share one temporal patch), and the rate is the smaller of a preset cap and token budget / (tokens per frame $\times$ total duration). Each episode is then encoded as a low-frame-rate MP4 at this rate, preserving within-episode temporal continuity while sampling longer series more sparsely; timestamps are recovered by the model's video processor from the container frame rate.

Because these models expose heterogeneous native video interfaces, the results should not be interpreted as comparisons under a fixed visual-input budget. Instead, they reflect end-to-end performance under each model's practical long-video interface.

5.2 Main Benchmark Results

Table 4 reveals a clear advantage of agentic and tool-augmented approaches on V-Horizon. **Agentic Skills achieves the best overall accuracy of 70.3%**, outperforming Qwen MM Plugins by 10.3 points (60.0%) and the strongest closed-source video model, Gemini 3.1 Pro, by 15.0 points (55.3%). More importantly, the gain is concentrated on questions that require explicit visual grounding: Agentic Skills reaches 65.6% on V+W, exceeding Gemini 3.1 Pro by 21.8 points (43.8%). It also obtains the highest accuracy on both V (60.4%) and W (70.9%). These results indicate that the advantage of agentic systems is not merely stronger language reasoning, but their ability to actively locate and inspect task-relevant visual evidence in long videos.

In contrast, conventional video models remain substantially stronger on text-dominant questions than on visually grounded ones. For example, Gemini 3.1 Pro achieves 79.1% on T+N but only 43.8% on V+W, while GPT-5.6 obtains 80.5% versus 41.1%, yielding gaps of 35.3 and 39.3 points, respectively. Increasing the visual input budget does not consistently resolve this limitation: across open-source models, using more frames or visual tokens yields mixed gains and can even degrade performance. This suggests that simply exposing a model to more of the video is insufficient for reliably identifying the evidence required by a question.

The frame- and token-selection baselines lead to a similar conclusion. Question-aware frame selection (CLIP and AKS) and visual-token compression (CDPruner and PruMerge) reduce the visual input presented to the model, yet none closes the performance gap to agentic approaches. On Qwen3-VL-8B, the strongest selection baseline, CDPruner, reaches only 34.7% overall accuracy; on InternVL3.5-8B, all selection variants remain below 30%. Taken together, these results suggest that long-video understanding requires more than increasing, pruning, or ranking visual inputs in a single pass. The key capability is to *adaptively decide what evidence is needed, where to retrieve it, and how to refine the search* as the reasoning process unfolds.

5.3 Finding 1: The Pseudo-Understanding Phenomenon

Table 4 reveals a systematic asymmetry between text-inferable and visually grounded capabilities. Across all evaluated configurations, pooled accuracy on text-inferable questions (T+N) is higher than that on visually grounded questions (V+W). We quantify this discrepancy using the *text–vision gap*, defined as

$$\Delta_{\mathrm{TV}} = \mathrm{Acc}_{T+N} - \mathrm{Acc}_{V+W} = \frac{c_T + c_N}{n_T + n_N} - \frac{c_V + c_W}{n_V + n_W}, \tag{1}$$

where c_k and n_k denote the number of correct answers and the total number of questions for capability $k \in \{T, N, V, W\}$, respectively. Positive and negative Δ_{TV} indicate higher accuracy on T+N and V+W, respectively. Balance is measured by proximity to zero, i.e., a smaller $|\Delta_{\mathrm{TV}}|$, rather than a smaller signed value. Equation 1 uses accuracy proportions; the tables report $100\Delta_{\mathrm{TV}}$ in percentage points. A near-zero gap can also arise when both groups have low accuracy: InternVL3.5-8B (240K VT) has a gap of only 1.9 points but an overall accuracy of 28.5%. The gap must therefore be interpreted jointly with both pooled accuracies and overall accuracy; it does not independently establish visual grounding or reliance on textual shortcuts.

This gap is especially pronounced among strong holistic video models. GPT-5.6 achieves 80.5% accuracy on T+N but only 41.1% on V+W, yielding the largest gap of 39.3 percentage points. Similarly, Gemini 3.1 Pro reaches 79.1% on T+N versus 43.8% on V+W, corresponding to a 35.3-point gap. These results show that strong aggregate performance can coexist with substantially weaker performance once explicit visual grounding is required.

We refer to this phenomenon as *pseudo-understanding*: a model may appear to understand a video well at the aggregate level while relying disproportionately on information that can be inferred from language, context, or prior knowledge rather than from the relevant visual evidence itself. The large Δ_{TV} values observed in high-performing holistic models indicate that improvements in language-supported reasoning do not necessarily translate into comparable gains in visually grounded ultra-long video understanding. Reporting Δ_{TV} alongside both pooled accuracies and aggregate accuracy characterizes the distribution of performance across capability groups; establishing the evidence used by a model requires controlled input ablations.

Table 5: **Per-axis comparison with Gemini 3.1 Pro.** Δ denotes the accuracy change of Agentic Skills over Gemini 3.1 Pro.

Model	T↑	N↑	V↑	W↑	T+N↑	V+W↑
Gemini 3.1 Pro	82.5	**76.3**	41.0	46.6	79.1	43.8
Agentic Skills (Ours)	**87.6**	73.7	**60.4**	**70.9**	**80.0**	**65.6**
Δ	**+5.1**	−2.6	**+19.4**	**+24.3**	**+0.9**	**+21.8**

Table 6: **Oracle retrieval with ground-truth evidence spans.** The baseline uses 512 uniformly sampled frames ("512 f"); the oracle rows additionally supply the ground-truth evidence frames. Overall and pooled accuracies are question-count-weighted.

Model	Setting	Overall↑	T↑	N↑	V↑	W↑	V+W↑
Gemini 3.1 Pro	512 f	58.3	87.8	82.7	43.5	47.3	45.4
	Oracle	**76.5**	**95.2**	**90.6**	**51.3**	**85.9**	**68.6**
Gemini 3 Flash	512 f	53.6	75.6	78.6	38.4	46.0	42.2
	Oracle	**75.5**	**92.1**	**86.0**	**53.6**	**84.4**	**69.0**

5.4 Finding 2: The Retrieval Bottleneck

Our analyses consistently identify *evidence retrieval* as a major bottleneck in ultra-long video understanding. Agentic Skills achieves 70.3% overall accuracy, outperforming Gemini 3.1 Pro (55.3%) by 15.0 percentage points. Importantly, this gain is concentrated on visually grounded questions rather than text-inferable ones. As shown in Table 5, the two systems perform similarly on pooled T+N accuracy (80.0% vs. 79.1%), whereas Agentic Skills improves substantially on V and W by 19.4 and 24.3 points, respectively, resulting in a 21.8-point gain on pooled V+W accuracy. This sharp asymmetry suggests that the primary benefit of the agentic pipeline lies in locating task-relevant moments and exposing the corresponding visual evidence for focused inspection.

Oracle retrieval provides more direct evidence for this hypothesis. When models are given frames from ground-truth evidence spans, thereby largely removing the temporal search problem, performance improves substantially on visually grounded questions (Table 6, whose baseline uses 512 uniformly sampled frames). For Gemini 3.1 Pro, V+W accuracy increases from 45.4% to 68.6%, with W improving particularly strongly from 47.3% to 85.9%; Gemini 3 Flash shows a similar pattern, with V+W rising from 42.2% to 69.0%. These large gains indicate that models already possess much of the required visual understanding; the limitation is that the relevant evidence is not reliably surfaced from the full video.

This conclusion is further supported by scaling and failure analysis. Within the Qwen family, scaling from Qwen3-VL-8B to Qwen3.6-35B yields a much smaller improvement than equipping the smaller backbone with Agentic Skills, suggesting that improving evidence access can outweigh increasing model capacity. Manual inspection of 100 Agentic Skills failures attributes 42% of errors to search misses, 31% to visual perception, and 27% to reasoning, indicating that reliably *finding* the relevant evidence is a larger obstacle than *interpreting* it once observed.

6 Conclusion

V-Horizon reveals a key limitation of current ultra-long video models: strong aggregate performance can mask weak visual grounding and reliance on text-inferable cues. We show that text-inferable and visually grounded capabilities must be evaluated separately, and that oracle retrieval yields substantial gains when relevant evidence is provided, identifying long-range evidence retrieval as the central bottleneck and motivating retrieval-centric, agentic systems.

Limitations. V-Horizon currently centers on television series and relies on offline preprocessing with a proprietary orchestrator, which may limit generalization to other long-video domains.

AI USE STATEMENT

AI tools were used to assist with language editing and proofreading, including improving grammar, clarity, and overall readability. The ideas, arguments, and final content remain my own, and I reviewed and approved all AI-assisted revisions.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

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

Yuntao Bai, Saurav Kadavath, Sandipan Kundu, Amanda Askell, Jackson Kernion, Andy Jones, Anna Chen, Anna Goldie, Azalia Mirhoseini, Cameron McKinnon, et al. Constitutional ai: Harmlessness from ai feedback. *arXiv preprint arXiv:2212.08073*, 2022.

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

Guo Chen, Yicheng Liu, Yifei Huang, Yuping He, Baoqi Pei, Jilan Xu, Yali Wang, Tong Lu, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. *arXiv preprint arXiv:2412.12075*, 2024a.

Guo Chen, Yicheng Liu, Yifei Huang, Baoqi Pei, Jilan Xu, Yuping He, Tong Lu, Yali Wang, and Limin Wang. Cg-bench: Clue-grounded question answering benchmark for long video understanding. In *International Conference on Learning Representations*, volume 2025, pp. 45647–45682, 2025.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216*, 4(5), 2024b.

Rohan Choudhury, Koichiro Niinuma, Kris M Kitani, and László A Jeni. Zero-shot video question answering with procedural programs. *arXiv preprint arXiv:2312.00937*, 2023.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *arXiv preprint arXiv:2406.14515*, 2024.

Kaituo Feng, Kaixiong Gong, Bohao Li, Zonghao Guo, Yibing Wang, Tianshuo Peng, Junfei Wu, Xiaoying Zhang, Benyou Wang, and Xiangyu Yue. Video-r1: Reinforcing video reasoning in mllms. *Advances in Neural Information Processing Systems*, 38:99114–99137, 2026.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118. IEEE, 2025.

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

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

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. In *International Conference on Learning Representations*, volume 2026, pp. 109089–109117, 2026.

Jingyang Lin, Jialian Wu, Jiang Liu, Ximeng Sun, Ze Wang, Xiaodong Yu, Jiebo Luo, Zicheng Liu, and Emad Barsoum. Videoseek: Long-horizon video agent with tool-guided seeking. *arXiv preprint arXiv:2603.20185*, 2026.

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

Qwen Team. Qwen-mm-plugins: Native multimodal plugins for qwen models. `https://github.com/QwenLM/Qwen-MM-Plugins`, 2026. GitHub repository.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *International conference on machine learning*, pp. 28492–28518. PMLR, 2023.

Bytedance Seed. Seed2. 0 model card: Towards intelligence frontier for real-world complexity. *arXiv preprint arXiv:2607.00248*, 2026.

Yuzhang Shang, Mu Cai, Bingxin Xu, Yong Jae Lee, and Yan Yan. Llava-prumerge: Adaptive token reduction for efficient large multimodal models. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 22857–22867. IEEE, 2025.

Dídac Surís, Sachit Menon, and Carl Vondrick. Vipergpt: Visual inference via python execution for reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 11888–11898, 2023.

Xi Tang, Jihao Qiu, Lingxi Xie, Yunjie Tian, Jianbin Jiao, and Qixiang Ye. Adaptive keyframe sampling for long video understanding. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 29118–29128. IEEE, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024a.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, et al. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*, 2025a.

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

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

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

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Shuhong Ye, Weikai Kong, Chenglin Yao, Jianfeng Ren, and Xudong Jiang. Video question answering using clip-guided visual-text attention. In *2023 IEEE International Conference on Image Processing (ICIP)*, pp. 81–85. IEEE, 2023.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025.

Qizhe Zhang, Mengzhen Liu, Lichen Li, Ming Lu, Yuan Zhang, Junwen Pan, Qi She, and Shanghang Zhang. Beyond attention or similarity: Maximizing conditional diversity for token pruning in mllms. *Advances in Neural Information Processing Systems*, 38:25438–25468, 2026a.

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

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

Bin Zhu, Bin Lin, Munan Ning, Yang Yan, Jiaxi Cui, WANG HongFa, Yatian Pang, Wenhao Jiang, Junwu Zhang, Zongwei Li, et al. Languagebind: Extending video-language pretraining to n-modality by language-based semantic alignment. In *International Conference on Learning Representations*, volume 2024, pp. 9588–9608, 2024.

A Reproducibility Statement

We plan to publicly release the V-Horizon benchmark, including the 660 curated questions, six answer options, answer keys, and temporally localized evidence spans, together with the preprocessing, question-generation, filtering, and evaluation code. Agentic Skills is built from publicly available components—BM25-based text search, LanguageBind embeddings for clip retrieval, and the Qwen3-VL-8B viewer—and we will release the orchestrator prompts and tool definitions. For proprietary models, we document the exact prompting templates, input-construction procedures (frame counts, contact-sheet layouts, and segment counts), and decoding settings used in our evaluations, so that the reported numbers can be independently verified given API access.

B Ethics Statement

V-Horizon is constructed from publicly available television series and is intended solely for research on video-language model evaluation. We do not redistribute the underlying video content; we release only questions, options, answer keys, and brief evidence spans that consist of short textual excerpts. The benchmark does not involve human subjects, personal data collection, or the generation of new media, and it is not designed for real-time surveillance or monitoring applications.

C Question Generation Pipeline Details

Stage 1: Event-Grounded Question Generation. Each video series is partitioned into 10-second intervals, with each interval aligned with timestamped subtitles and dense visual captions. Based on this temporally aligned evidence, we construct axis-specific prompts to systematically probe four complementary reasoning dimensions:

- **Temporal (T):** Tests temporal ordering, event timing, and duration across the video.
- **Narrative (N):** Evaluates causal and narrative dependencies across events and episodes.
- **Visual (V):** Targets fine-grained visual details that cannot be reliably inferred from dialogue or subtitles alone.
- **Locating (W):** Requires identifying when and where a particular scene, action, or event takes place.

Each candidate question is generated from and grounded in timestamped evidence from the source videos, ensuring that its correct answer can be traced to a verifiable evidence span.

Stage 2: Distractor Synthesis. For each question, we generate five plausible distractors using three targeted confusion strategies: (1) *temporal confusion*, which replaces the correct event with an authentic event occurring at an incorrect timestamp; (2) *visual confusion*, which substitutes the target entity, object, attribute, or action with a semantically similar but incorrect alternative; and (3) *narrative confusion*, which introduces plausible but unsupported causal relations, event interpretations, or character motivations. These strategies encourage distractors that remain semantically coherent and contextually plausible while requiring precise evidence grounding to distinguish them from the correct answer.

Stage 3: Blind-Test Filtering. To reduce linguistic leakage and characterize the modality dependence of each question, we audit candidate questions using GPT-5.6 under three progressively richer evidence conditions: (i) the question and answer choices only; (ii) the question and answer choices together with subtitle transcripts; and (iii) the question and answer choices together with both subtitle and visual context. Questions that can be reliably answered under the question-only condition are discarded, as they may admit linguistic shortcuts or rely on parametric knowledge rather than evidence from the video. Among the remaining questions, performance under the subtitle-only condition is used to distinguish instances that can be resolved primarily from textual evidence from those that require additional visual grounding. This filtering procedure enables the benchmark to retain a controlled mixture of text-inferable and visually grounded questions while suppressing unintended answer leakage.

Stage 4: Difficulty Calibration. The 937 candidates retained after the preceding stages are further evaluated using 15 baseline models to identify trivial instances and reduce ceiling effects. We remove questions answered correctly by all 15 baselines, pruning 277 instances. After the full generation, filtering, and calibration pipeline, **V-Horizon contains 660 curated six-way multiple-choice questions**. The resulting benchmark naturally contains a higher proportion of visually grounded questions than text-inferable ones, reflecting the emphasis on evaluating long-horizon visual understanding beyond information recoverable from subtitles alone.

D Benchmark Quality Validation

Table 7: **Blind-test accuracy (%) of GPT-5.6 under progressively richer input conditions.** The results characterize the extent to which each capability axis can be resolved from textual versus visual evidence.

Input Condition	T	N	V	W
Question + options only	22.7	23.5	17.3	16.7
+ Subtitle transcripts	60.2	62.5	21.3	18.9
+ Subtitles & sampled frames	64.7	75.6	29.3	24.3

As shown in Table 7, accuracy under the question-only condition ranges from 16.7% to 23.5%, remaining near or only modestly above the 16.7% random-chance accuracy of a six-way multiple-choice task. This result indicates limited answerability from parametric knowledge or superficial linguistic cues alone.

Providing subtitle transcripts leads to a pronounced improvement on the Temporal and Narrative axes: accuracy increases from 22.7% to 60.2% (**+37.5** points) on T and from 23.5% to 62.5% (**+39.0** points) on N. In contrast, the corresponding gains on Visual and Locating questions are only **+4.0** and **+2.2** points, respectively. This pronounced separation supports the intended modality structure of V-Horizon: T and N are predominantly supported by textual evidence, whereas V and W require substantially more information beyond subtitles.

Adding uniformly sampled frames provides further evidence across all four axes. Performance on V and W improves by **+8.0** and **+5.4** points over the subtitle-only condition, confirming that visual observations provide information unavailable in the transcripts. However, absolute accuracy remains low on both axes, reaching only 29.3% on V and 24.3% on W. These results suggest that sparse uniform sampling exposes only a fraction of the visual evidence needed to resolve visually grounded questions, highlighting the importance of effective evidence localization and retrieval for long-horizon video understanding.

Table 8: **Average tool invocations per query across capability axes.** Runtime visual calls aggregate clip retrieval and VLM viewer usage.

Tool	T	N	V	W
Text Search	2.8	3.1	1.9	3.4
Clip Retrieval	0.3	0.2	1.4	0.8
VLM Viewer	0.1	0.1	1.2	0.5
Runtime Visual Calls	0.4	0.3	**2.6**	**1.3**
Total Calls	**3.2**	**3.4**	**4.5**	**4.7**

E Tool Usage Patterns

Table 8 summarizes the agent's tool usage across the four capability axes. The agent exhibits a clear modality-dependent retrieval pattern. For the more text-inferable Temporal and Narrative questions, it relies predominantly on text search, averaging 2.8 and 3.1 calls per query, respectively, while invoking runtime visual tools only occasionally. In contrast, visually grounded questions induce

Table 9: **Ablation of runtime visual tools in Agentic Skills.** The ablated setting disables clip retrieval and the VLM viewer while retaining text retrieval over the precomputed multimodal index, including dense visual captions.

Configuration	Overall↑	T↑	N↑	V↑	W↑
Full Agentic Skills	**70.3**	**87.6**	73.7	**60.4**	**70.9**
w/o Runtime Visual Tools	69.6	87.1	**76.8**	59.5	68.1
Δ (Full − Ablated)	**+0.7**	+0.5	−3.1	+0.9	**+2.8**

substantially more visual processing. For Visual questions, clip retrieval and the VLM viewer are invoked 1.4 and 1.2 times per query on average, respectively; for Locating questions, the corresponding frequencies are 0.8 and 0.5 calls per query.

These patterns indicate that the agent adapts its retrieval strategy to the evidence requirements of each query rather than applying a fixed sequence of tools. In particular, runtime visual tools are invoked more frequently when textual retrieval alone is less likely to provide sufficient evidence. Across the four capability axes, the agent makes an average of 3.2–4.7 tool calls per instance, with an end-to-end response latency of 38–62 seconds. This selective tool-use behavior allows the system to allocate additional visual computation primarily to queries for which it is most useful.

F Tool Ablation

Interestingly, the performance gains of the agent do not depend strongly on runtime visual tool invocation. As shown in Table 9, disabling both CLIP-based clip retrieval and the VLM viewer, while retaining text search over the precomputed multimodal index, reduces overall accuracy from 70.3% to 69.6%, a drop of 0.7 percentage points. This result suggests that dense visual captions encode a substantial portion of the task-relevant visual content into a representation that can be efficiently accessed through text retrieval.

Importantly, this ablation does not remove visual information altogether: although runtime access to raw visual content is disabled, the searchable index still contains visual evidence distilled into dense captions during offline preprocessing. The relatively small performance gap therefore suggests that much of the benefit of Agentic Skills arises from *structured multimodal indexing and targeted evidence retrieval*, while direct runtime visual inspection provides a complementary gain for cases that cannot be resolved from the indexed textual representations alone. This decomposition offers a computationally efficient path toward ultra-long-horizon video understanding, where most queries can be served through lightweight retrieval and additional visual computation is reserved for evidence-demanding cases.