VidOmni-Bench: A Benchmark for Fine-Grained Video Understanding via Spatio-Temporal Event Verification across Complexity and Duration

Changbeen Kim[1] **Junwon Chang**[1] **Kipyo Kim**[2] **Risa Shinoda**[3]
Kuniaki Saito[4] **Donghyun Kim**[1]*

[1]Korea University [2]Sungkyunkwan University
[3]The University of Tokyo [4]OMRON SINIC X Corporation
https://github.com/changbeendata/VidOmni-Bench

Abstract

While Video Large Language Models (Video-LLMs) have recently demonstrated strong performance, reliably evaluating their fine-grained video understanding remains challenging. Existing benchmarks often rely on question answering or ground-truth caption matching, where models may succeed through superficial cues and incomplete annotations. To this end, we introduce VidOmni-Bench, a benchmark that requires models to verify whether each event in dense video captions is supported by the video. VidOmni-Bench consists of 500 videos spanning five complexity types and diverse durations from 4 seconds to 90 minutes. After collecting videos along these axes, we use diverse Video-LLMs to generate dense captions and obtain human-verified sentence-level labels, where sentences containing incorrect events serve as hard negatives for evaluation. Our experiments on VidOmni-Bench reveal three key findings: (i) Video-LLMs frequently generate hallucinated descriptions in dense video captioning; (ii) they also struggle as verifiers, failing to reliably detect plausible but incorrect event descriptions; and (iii) model weaknesses vary across video complexity and duration, revealing diverse, model-specific bottlenecks in current Video-LLMs.

1 Introduction

Video Large Language Models (Video-LLMs) (Tang et al., 2026; Bai et al., 2025; Wang et al., 2025b; Yang et al., 2025b) have demonstrated strong capabilities in video understanding tasks such as question answering and video retrieval. However, reliably evaluating their fine-grained video understanding remains challenging (Kong et al., 2025; Cai et al., 2024), as videos vary widely in duration and visual dynamics. A comprehensive evaluation should therefore assess whether models can interpret diverse video evidence across different temporal scales and visual aspects, including object states, actions, background changes, and scene transitions.

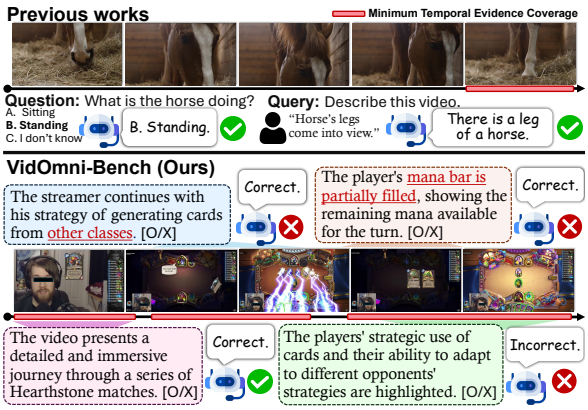

Figure 1: **Top:** Previous benchmarks typically rely on MCQA or sparse GT captions, allowing models to succeed with limited temporal evidence. **Bottom:** In VidOmni-Bench, models must verify each event in dense captions against the video, enabling a more rigorous assessment of fine-grained video understanding.

Recent temporal and long-form video benchmarks mostly rely on Multiple-Choice Question Answering (MCQA) (Fu et al., 2025; Fang et al., 2024; Tan et al., 2025) or Ground-Truth (GT) caption matching (Rawal et al., 2025; Liu et al., 2025; Lee et al., 2025). MCQA-based evaluations may allow models to rely on localized evidence, or textual shortcuts without temporal reasoning (Feng et al., 2025). Caption-matching-based benchmarks can assess video understanding more rigorously, but their focus on relatively short videos and human-written GT captions limits scalable dense event-level coverage across diverse video durations. Consequently, existing benchmarks do not directly evaluate whether Video-LLMs can understand fine-grained video events by integrating broad temporal evidence across diverse durations.

To address these gaps, we introduce VidOmni-Bench, a sentence-wise event verification bench-

*Corresponding author: d_kim@korea.ac.kr

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 2459–2477
October 24-29, 2026 ©2026 Association for Computational Linguistics

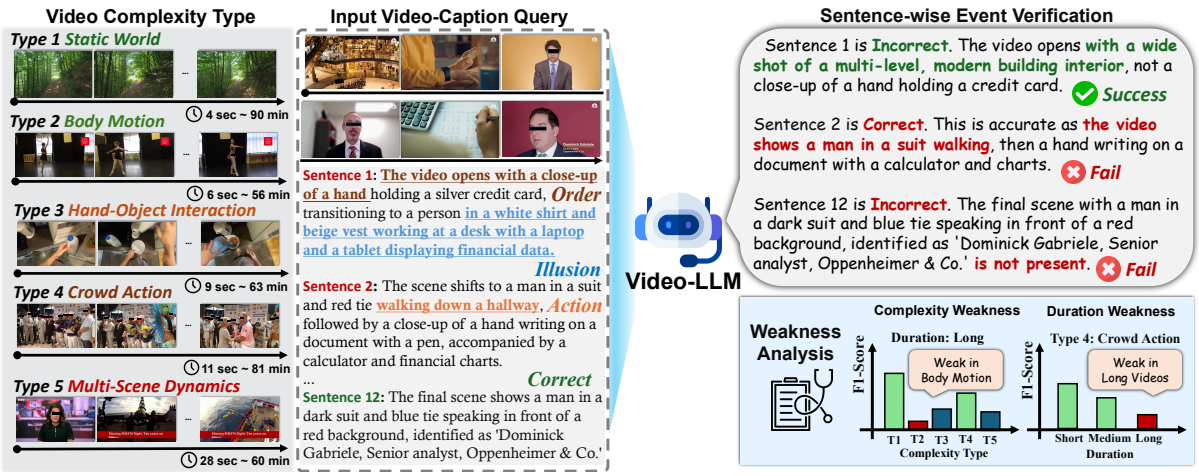

Figure 2: We introduce VidOmni-Bench, a novel benchmark designed to evaluate how well Video-LLMs distinguish between correct and incorrect events across five types of spatio-temporal complexity and diverse video durations. This structured framework enables a rigorous assessment of whether Video-LLMs can truly ground their reasoning in the visual evidence across diverse scales.

mark where a model verifies whether each sentence in dense video captions is supported by the video. As shown in Figure 2, VidOmni-Bench consists of 500 videos curated across five complexity types and diverse durations from 4 seconds to 90 minutes. For each video, we use a diverse set of Video-LLMs to generate dense captions, and human annotators verify each sentence to produce sentence-level labels. Hallucinated or incorrect sentences are then used as hard negatives for evaluation, resulting in a benchmark designed to assess event verification across diverse types of video complexity and duration. As shown in Table 2, VidOmni-Bench requires models to inspect substantially broader portions of the video than existing benchmarks, indicating that VidOmni-Bench cannot be solved through superficial cues. Furthermore, Figure 7 shows that the captions in VidOmni-Bench provide dense and complex event descriptions across the five complexity types. These results support VidOmni-Bench as a challenging benchmark for fine-grained video understanding.

Our experiments on VidOmni-Bench reveal several critical limitations of current Video-LLMs. First, Video-LLMs frequently produce hallucinated descriptions during dense video captioning, indicating unreliable video-grounded generation. Second, they also struggle as verifiers, failing to reliably detect plausible but incorrect event descriptions within these captions. Third, model weaknesses vary substantially across video complexity and duration, revealing diverse, model-specific bottlenecks rather than a single uniform failure mode. These findings indicate that current Video-LLMs still lack robust, temporally grounded event understanding needed to faithfully interpret dense and diverse video content.

2 Related Works

Video Large Language Models. Following the success of Large Language Models (LLMs) (Yang et al., 2025a; Grattafiori et al., 2024), multimodal understanding has expanded from static images (Liu et al., 2023; Chen et al., 2024b) to dynamic videos (Li et al., 2024, 2025; Chen et al., 2024a; Bai et al., 2025). Recent Video-LLMs can generate detailed video descriptions, but their outputs often contain subtle hallucinations (Gao et al., 2025). We use these generated descriptions for sentence-wise event verification, with human annotators labeling hallucinated sentences as hard negatives.

Video Understanding Benchmarks. Recent benchmarks have advanced the evaluation of temporal and long-form video understanding (Ben Ami et al., 2026; Ha et al., 2026; Li et al., 2026; Cai et al., 2024; Kong et al., 2025), but many remain based on MCQA formats. Such positive-selection formats can be affected by language priors or partial visual evidence (Feng et al., 2025), and provide limited evidence of whether models can reject plausible but incorrect events. VidOmni-Bench instead evaluates dense event verification beyond sparse question-answer pairs.

Hallucination in Video Understanding. Several studies evaluate video captioning and hallucination

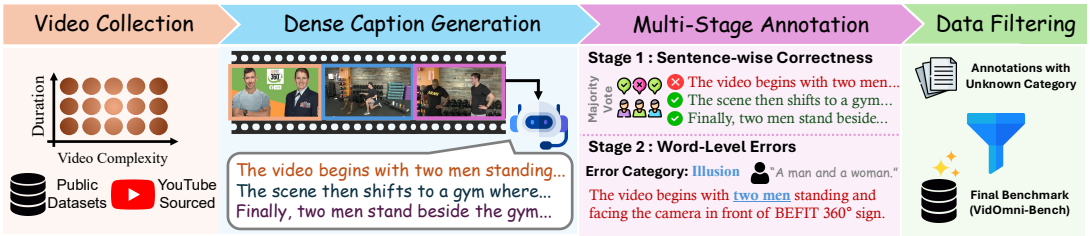

Figure 3: **Construction pipeline of VidOmni-Bench.** We collect videos from public datasets and YouTube along with two axes. Then, we utilize diverse Video-LLMs for dense caption generation. These video and dense caption pairs then undergo a multi-stage annotation phase for sentence-wise error (hallucination) labeling, where annotators label sentence-level errors to construct hard negatives. Finally, the data filtering stage for a robust evaluation.

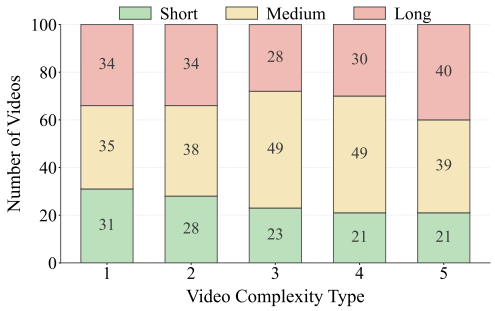

Figure 4: Our dataset consists of 500 videos, with exactly 100 videos allocated to each of the five complexity types. Within each type, we maintain a balanced distribution across Short, Medium, and Long duration.

detection by comparing model outputs with human-written GT captions (Liu et al., 2025; Rawal et al., 2025; Lee et al., 2025). While useful for assessing factual consistency, such sparse references limit dense event-level coverage across diverse domains and durations. Motivated by Saito et al. (2026), VidOmni-Bench uses Video-LLM-generated dense descriptions and human-verified negatives to analyze hallucination patterns across video complexity and duration.

3 VidOmni-Bench

To evaluate fine-grained understanding of Video-LLMs, we introduce a sentence-wise event verification task. This task requires models to act as verifiers, determining both the factual occurrence of described events and their temporal consistency within the video. To provide a granular diagnosis of model capabilities, we construct VidOmni-Bench by disentangling spatio-temporal complexity from duration. Rather than relying on human-written captions, we generate dense captions using Video-LLMs and verify them at the sentence level. The following sections define the task, describe the video collection and annotation process, and present statistical analysis. The construction process of VidOmni-Bench is illustrated in Figure 3.

3.1 Task Definition

Sentence-wise event verification evaluates whether each sentence in a video caption is factually supported by the video. Formally, given a video V and an n-sentence caption $C = \{s_1, s_2, \ldots, s_n\}$, the goal is to predict binary labels $y = \{y_1, y_2, \ldots, y_n\}$. For each sentence s_i, we define $y_i \in \{0, 1\}$ as:

$$y_i = \begin{cases} 0, & \text{if } s_i \text{ is factually supported by } V, \\ 1, & \text{otherwise.} \end{cases} \quad (1)$$

Thus, a Video-LLM must verify each sentence against the video evidence and determine whether it contains hallucinated or incorrect content.

3.2 Video Collection

To enable a granular diagnosis of Video-LLMs, we curate videos by controlling spatio-temporal complexity and video duration as two separate axes. We collect videos from diverse public datasets, covering egocentric actions (Grauman et al., 2022; Damen et al., 2018; Sener et al., 2022), anomaly events (Sultani et al., 2018; Wu et al., 2020), and long-form content (Caba Heilbron et al., 2015; Fu et al., 2025). We further supplement these sources with targeted YouTube crawling to fill underrepresented combinations of duration and complexity. For crawled videos, we only retain content released under Creative Commons licenses.

Our collection is stratified into five predefined spatio-temporal complexity types: (1) Static World, (2) Body Motion, (3) Hand-Object Interaction, (4) Crowd Action, and (5) Multi-Scene Dynamics. As shown in Figure 4, each complexity type is balanced across three duration brackets: short ($\leq$ 3 min), medium (4–30 min), and long ($\geq$ 30 min). This design enables us to analyze the effects of complexity and duration separately. The resulting

VidOmni-Bench contains 500 videos, 433 of which include audio tracks. Detailed definitions of complexity types and data sources are provided in the Appendix B.

3.3 Annotation Pipeline

To overcome the limited coverage of human-written GT captions, we use modern Video-LLMs to generate dense descriptions and repurpose their human-verified hallucinations as hard negatives. We use five Video-LLMs with diverse architectures and parameter scales: InternVL3.5 (Wang et al., 2025b) (38B), Qwen3-VL (Bai et al., 2025) (32B), video-SALMONN 2+ (Tang et al., 2025) (72B), Keye-VL 1.5 (Yang et al., 2025b) (8B), and Gemini 3 Flash (Google, 2025). For each video, one of five specialized chat templates (e.g., "Describe the full content of the video in detail.") is randomly assigned to each model, producing 2,500 video-caption pairs in total. Using multiple models and prompts diversifies the generated descriptions and reduces dependence on the output of a single model.

These raw captions serve as the basis for a two-stage human annotation process. In the first stage, five independent annotators verify the factual correctness of each sentence and label it as Correct, Incorrect, or Unknown. Final labels are determined by majority vote, and Unknown cases are excluded to maintain reliable evaluation.

In the second stage, we perform word-level error annotation for sentences labeled as Incorrect. Annotators identify the incorrect words within each sentence and assign each error to one of ten categories: five related to static-frame understanding (Object, Relation, Location, Direction, and Illusion) and five related to temporal understanding (Action, Order, Causality, Frequency, and State Change). The resulting human-verified incorrect sentences serve as hard negatives for sentence-wise event verification. Details for dense captioning, and annotation process are provided in the Appendix B.

3.4 Data Statistics and Analysis

Basic statistics. Table 1 summarizes VidOmni-Bench, which contains 64K annotated sentences. We report the Video Incorrect Rate, defined as the proportion of video-caption pairs in which the generated caption contains at least one hallucinated sentence, along with the average sentence-level Incorrect Rate per caption. The results show that

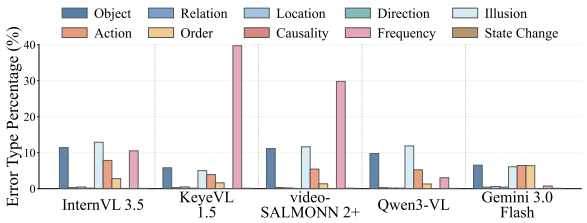

Figure 5: **Word-level error statistics per category.** The consistent high frequency of temporal errors in dense video captioning highlights a fundamental bottleneck in video understanding.

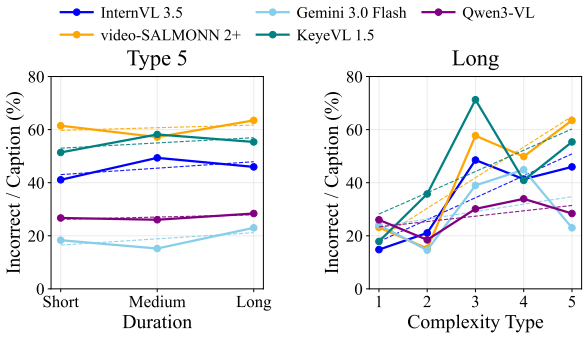

Figure 6: **Effect of video complexity and duration on hallucination.** Plots show the average sentence-level Incorrect Rate per caption. **Left**: across durations for Type 5 videos. **Right**: across complexity types within Long videos. Dashed lines indicate overall trends.

Video-LLMs frequently generate hallucinated captions, with Incorrect rates remaining high across all models, thereby providing a sufficient pool of hard negatives for sentence-wise event verification.

Error category distribution. As shown in Figure 5, temporal errors such as Action, Order, and Frequency account for a substantial portion of hallucinations, suggesting that current Video-LLMs struggle with fine-grained temporal reasoning. Figure 2 further shows that a sentence can appear plausible in isolation but become Incorrect when verified against the broader video context.

Effect of complexity and duration on dense captioning. By controlling video complexity and duration, we analyze hallucination patterns in dense captioning (Figure 6). The sentence-level Incorrect Rate per caption generally increases with complexity type within each duration bracket, while Type 5 remains consistently high across all durations. This suggests that hallucinations are driven not only by video length, but also by the difficulty of reasoning over complex event dynamics.

Caption characteristics differ across complexity types. We analyze how caption properties and

	KeyeVL 1.5	InternVL 3.5	Qwen3-VL	video-SALMONN 2+	Gemini 3
Params.	8B	38B	32B	72B	N/A
Total Sentences	18974	5542	11868	18974	9485
Video Incorrect Rate (%)	92.6	**85.6**	94.0	92.2	91.1
Incorrect / Caption (%)	40.6	39.5	26.5	44.9	**24.1**

Table 1: **Statistics of incorrect sentences in generated captions.** VidOmni-Bench provides a large set of annotated incorrect sentences for evaluating sentence-wise event verification. Sentences labeled as Unknown are excluded.

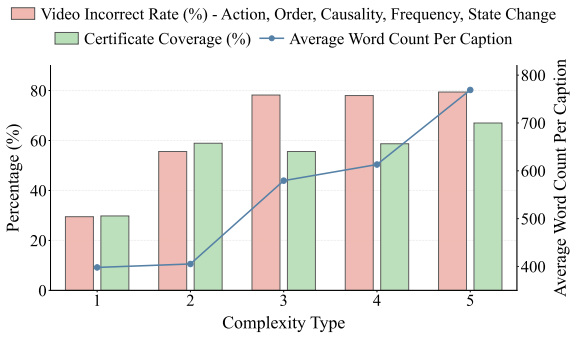

Figure 7: **Caption characteristics across complexity types.** Higher-complexity videos lead to longer captions and a higher Video Incorrect Rate for temporal error categories, while also requiring broader video evidence in sentence-wise event verification of VidOmni-Bench.

verification requirements vary across the five complexity types in VidOmni-Bench. Figure 7 shows that higher complexity is associated with longer captions and higher Video Incorrect Rates over temporal error categories(Action, Order, Causality, Frequency, and State Change). We further report Certificate Coverage following Video-MME (Fu et al., 2025), showing that sentence-wise event verification generally requires broader video evidence as complexity increases. Details for calculating Certificate Coverage are provided in the Appendix D.

Comparison with existing benchmarks. Figure 1 and Table 2 position VidOmni-Bench among existing video understanding benchmarks. Moving beyond MCQA-based evaluation (Fu et al., 2025; Ben Ami et al., 2026; Li et al., 2026), VidOmni-Bench performs sentence-wise event verification over dense descriptions. Compared with benchmarks based on human-written GT caption matching (Rawal et al., 2025; Lee et al., 2025; Liu et al., 2025; Ma et al., 2025), VidOmni-Bench spans broader video durations. Its higher Certificate Coverage further shows that verification requires broader video evidence, while the Appendix D shows that VidOmni-Bench captions are more complex than those in prior benchmarks.

Benchmark	Evaluation Task	Duration	Cert. Cov. (%)
TimeBlind (Li et al., 2026)	MCQA	$\leq$ 25s	32.3
NOAH (Lee et al., 2025)	GT Matching / MCQA	$\leq$ 12m	30.8
ARGUS (Rawal et al., 2025)	GT Matching	$\leq$ 3m	25.0
HERBench (Ben Ami et al., 2026)	MCQA	$\leq$ 6m	24.8
Video-MME (Fu et al., 2025)	MCQA	$\leq$ 60m	3.4
CAPAbility (Liu et al., 2025)	GT Matching / MCQA	$\leq$ 50s	2.0
VideoEval-Pro (Ma et al., 2025)	GT Matching	$\geq$ 30m	0.9
VidOmni-Bench (Ours)	**Sentence Verification**	$\leq$ **90m**	**54.0**

Table 2: **Comparison with other benchmarks in task design and temporal evidence.** VidOmni-Bench evaluates sentence-wise event verification over longer videos and requires broader temporal evidence, indicated by higher Certificate Coverage (Cert. Cov.).

4 Experiments

In this section, we evaluate a wide range of Video-LLMs on VidOmni-Bench to assess their fine-grained video understanding. We first present quantitative results, followed by a detailed analysis of model performance across different dimensions. In summary, our analysis yields five key findings: (i) VidOmni-Bench reveals verification weaknesses not captured by existing benchmarks, showing that models often fail to consistently identify hallucinated sentences; (ii) scaling model size mainly improves verification of static-frame-related errors, whereas temporal ordering and causal reasoning remain major bottlenecks; (iii) current Video-LLMs exhibit a distinct self-preference bias, showing unreliable verification behavior when judging their own generated sentences; (iv) audio often provides verification-relevant information complementary to visual evidence, and models that effectively incorporate these cues achieve stronger performance; and (v) model vulnerabilities are multidimensional, with performance depending on model-specific sensitivity to both complexity and duration.

4.1 Experimental Setup

Benchmark coverage and evaluation protocols. We evaluate Video-LLMs on VidOmni-Bench together with three existing video understanding benchmarks: Video-MME (Fu et al., 2025), Video-

Video-LLM	Modality	Video-MME	Video-MMMU	LVBench	VidOmni-Bench		
		QA-Acc. (%)	QA-Acc. (%)	QA-Acc. (%)	Precision (%)	Recall (%)	F1-Score (%)
Open-Source Models							
Keye-VL 1.5 8B	v	73.0	66.0	35.1†	19.1	8.4	9.6
InternVL 3.5 2B	v	58.4	41.8†	30.5†	6.0	1.8	2.3
InternVL 3.5 8B	v	66.0	48.9†	34.2†	26.4	13.9	14.7
InternVL 3.5 14B	v	67.9	54.8†	35.1†	32.3	18.1	19.4
InternVL 3.5 38B	v	70.9	57.9†	38.1†	25.0	12.7	14.1
Qwen3-VL 2B	v	61.9	41.9	47.4	5.5	4.2	3.6
Qwen3-VL 4B	v	69.3	56.2	56.2	22.9	11.7	12.7
Qwen3-VL 8B	v	71.4	65.3	58.0	25.9	13.6	14.7
Qwen3-VL 32B	v	76.6	71.9	63.8	35.4	19.6	21.7
video-SALMONN 2+ 7B	v	67.8†	46.0†	49.7	1.3	1.1	1.1
video-SALMONN 2+ 7B	v, a	73.4	50.2†	47.8†	0.1	0.2	0.1
video-SALMONN 2+ 72B	v	74.0†	62.0†	55.5	8.3	5.4	5.8
video-SALMONN 2+ 72B	v, a	79.7	60.8†	55.2†	5.7	3.4	3.7
Closed-Source Models							
Gemini 3 Flash	v	73.2†	80.3†	74.3†	54.6	37.2	39.8
Gemini 3 Flash	v, a	86.1†	75.3†	76.9†	**54.7**	**39.9**	**41.7**
GPT 5 mini	v	71.0	56.7	57.8†	39.3	33.2	28.9

Table 3: **Main results on VidOmni-Bench.** We report verification Precision, Recall, and F1-Score for VidOmni-Bench, alongside QA Accuracy (QA-Acc) for Video-MME, Video-MMMU, and LVBench. Models are evaluated under video-only (v) or video+audio (v, a) settings. † denotes reproduced performance.

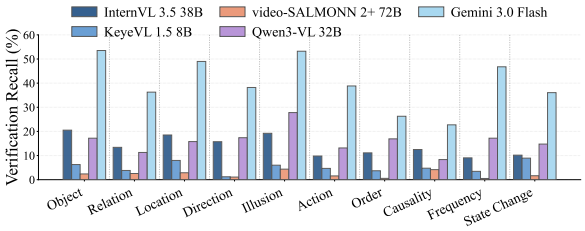

Figure 8: **Performance of Video-LLMs across error categories.** This visualization highlights the relative strengths and weaknesses of current Video-LLMs in detecting specific static and temporal discrepancies.

MMMU (Hu et al., 2026), and LVBench (Wang et al., 2025a). Since these benchmarks are formulated as Question Answering (QA) tasks, we report QA accuracy following their original evaluation protocols. In contrast, VidOmni-Bench is formulated as a sentence-wise event verification task, where sentences labeled Incorrect are treated as the detection target. For the i-th video-caption pair ($i = 1, \ldots, N$), we compute Precision, Recall, and F1-Score over all sentences within the pair:

$$\text{Precision}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FP}_i}, \quad \text{Recall}_i = \frac{\text{TP}_i}{\text{TP}_i + \text{FN}_i},$$

$$\text{F1-Score}_i = \frac{2 \cdot \text{Precision}_i \cdot \text{Recall}_i}{\text{Precision}_i + \text{Recall}_i}. \quad (2)$$

We then report the averaged scores over all N pairs.

Evaluated models. To analyze how model characteristics affect verification performance, our model selection in evaluation is guided by three criteria. First, we include models with varying parameter scales to examine the impact of model capacity. Second, for audio-capable models, we evaluate both video-only and video+audio settings to assess the contribution of audio signals to event verification. Third, we evaluate powerful closed-source models, including GPT 5 mini (Singh et al., 2026), as strong reference points for current model performance. Detailed model configurations and the evaluation prompt are provided in the Appendix C.

4.2 Overview of Results

Table 3 presents the quantitative evaluation of diverse Video-LLMs on VidOmni-Bench, together with their performance on existing benchmarks. Qualitative examples of Video-LLM outputs are shown in Figure 9.

VidOmni-Bench enables rigorous evaluation beyond existing benchmarks. Our sentence-wise event verification task effectively differentiates Video-LLM performance. High leaderboard scores on existing benchmarks such as Video-MME (Fu et al., 2025), Video-MMMU (Hu et al., 2026), and LVBench (Wang et al., 2025a) do not necessarily translate to strong performance on VidOmni-

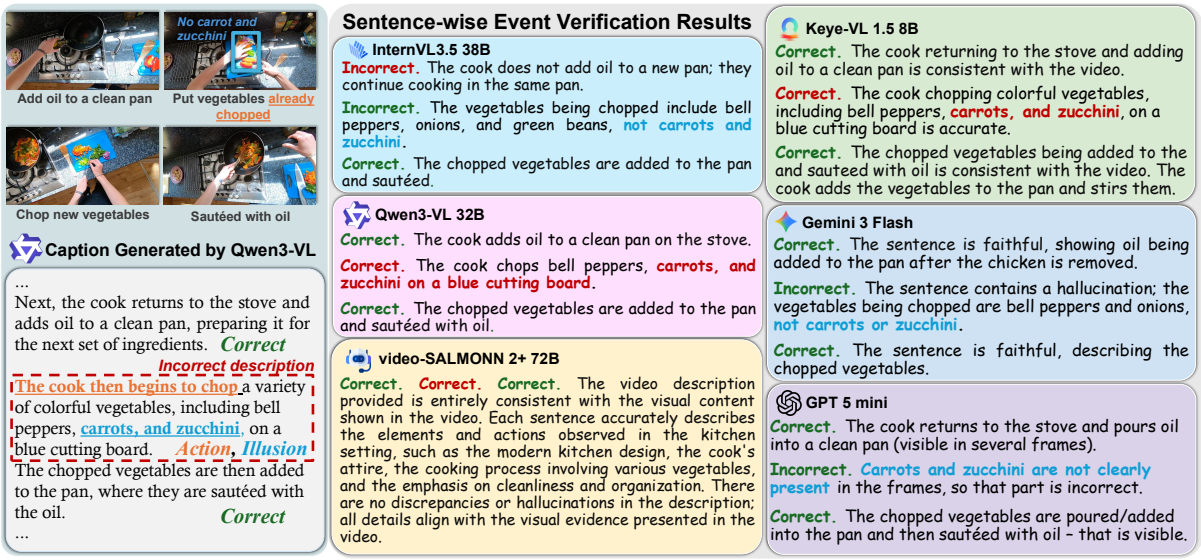

Figure 9: **Qualitative examples of sentence-wise event verification on a Qwen3-VL caption.** Video-LLMs may produce the same prediction while relying on different reasoning, revealing varying degrees of video understanding.

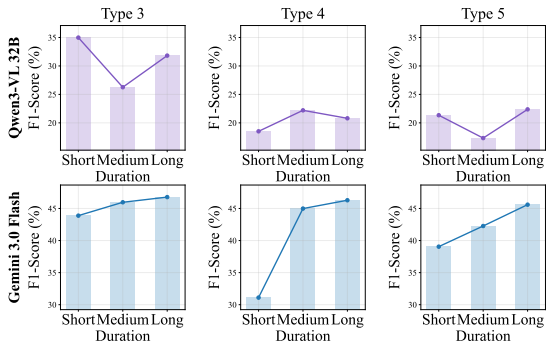

Figure 10: **F1-Score variations by video duration under fixed complexity.** Each column represents a specific complexity type, illustrating how performance changes across Short, Medium, and Long durations.

Bench. We further observe a significant gap between Precision and Recall across most models, indicating that they often fail to consistently identify hallucinated sentences. These findings show that VidOmni-Bench provides a more challenging testbed for fine-grained video understanding than existing evaluation benchmarks.

Performance comparison. Gemini 3 Flash shows the highest overall performance when using both video and audio inputs, serving as the strongest reference point among evaluated settings. Among open-source models, Qwen3-VL 32B performs best, while the InternVL3.5 family also shows competitive results. However, Figure 8 shows that InternVL3.5 38B is particularly strong at verifying static-frame-related hallucinations but weaker on temporal errors. This suggests that strong frame-level understanding does not necessarily translate to robust temporal reasoning.

Examples of Video-LLMs' output. Figure 9 shows sentence-wise event verification results from multiple Video-LLMs on a Qwen3-VL-generated caption. Each model outputs a sentence-level prediction together with reasoning, allowing us to qualitatively compare not only the final judgments but also the evidence used to support them.

Effect of scaling model size on event verification. Performance generally improves with model scale within the same family. Larger models achieve higher F1-Scores, suggesting that increased parameter capacity enhances their ability to detect subtle discrepancies between hard negative sentences and the corresponding video evidence.

Effect of audio modality on event verification. As shown in Table 3, adding audio consistently decreases the F1-Scores within the video-SALMONN 2+ family, indicating that audio may introduce cross-modal noise when it is not properly grounded with visual evidence. In contrast, Gemini 3 benefits from audio input, where audio often provides verification-relevant cues complementary to visual information. This suggests that audio can improve sentence-wise event verification when models effectively identify and integrate audio evidence relevant to each caption sentence.

4.3 Detailed Analysis

Success cases of event verification. As illustrated in Figure 9, successful verification is often tightly

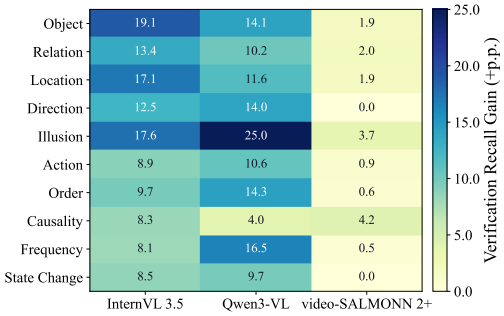

Figure 11: **Recall gain by error category as model size increases.** The heatmap shows the improvement in Recall for InternVL3.5, Qwen3-VL, and video-SALMONN 2+ when scaling up parameters.

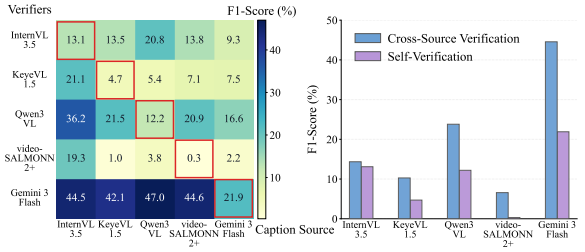

Figure 12: **Self-preference bias in self-verification. Left:** The heatmap shows F1-Scores for sentence-wise verification, with the y-axis representing the evaluated models and the x-axis representing the caption source models. **Right:** The bar chart compares the performance on cross-source verification versus self-verification.

coupled with factually grounded reasoning. This indicates that the model's output is based on an integrated understanding of the video content.

Failure cases of event verification. We observe two primary failure patterns. First, models may predict an Incorrect sentence as Correct by overlooking its mismatch with the video evidence. Second, even correct predictions can be accompanied by incomplete or factually flawed explanations. These cases highlight the need to examine both predictions and explanations when assessing whether verification is grounded in the video evidence.

Video-LLMs have their own weak complexity and duration. Figure 10 shows that performance trends across complexity and duration differ by model. For example, Gemini 3 Flash exhibits a consistent F1-Score decline as duration shortens within high-complexity types, while Qwen3-VL shows different vulnerabilities depending on the complexity type. These results demonstrate that VidOmni-Bench diagnoses multidimensional, model-specific weaknesses beyond a single difficulty axis. Further analyses are provided in the Appendix C.

Category-wise performance gains from model size scaling. To examine where performance improvements arise as Video-LLMs scale, we measure Recall gains across error categories when increasing model size for InternVL3.5 (2B to 38B), Qwen3-VL (2B to 32B), and video-SALMONN 2+ (7B to 72B). Figure 11 shows that scaling consistently improves Recall more for static-frame-related errors than for temporal-related errors. These results suggest that increasing model size within the same family mainly improves verification of static visual attributes and spatial relations, whereas temporal errors remain challenging.

Video-LLMs struggle to detect their own hallucinations. As shown in Figure 12, Video-LLMs perform worse on their own captions than on captions generated by other models. This trend is clear even for Gemini 3 Flash, the best-performing model. Its F1-Score exceeds 40 on captions from other models but drops by nearly half on its own captions. This suggests a strong self-preference bias, where models tend to judge their own generated sentences as Correct. Further analysis in the Appendix C shows that ensemble-based verification can mitigate this bias. These results indicate that self-evaluation remains a significant challenge for current Video-LLMs.

Ensemble can mitigate self-preference bias in self-evaluation. As shown in Table 4, cross-model ensemble substantially improves F1-Scores under self-evaluation settings. In this strategy, a generated sentence is marked as Incorrect if either the captioning model or an external verifier predicts it as Incorrect. The consistent gains suggest that self-evaluation failures are largely caused by missed hallucinations, and that external model verification can partially mitigate this self-bias.

Additional analysis in the Appendix. The Appendix D provides additional robustness analyses for VidOmni-Bench. Increasing the input frames does not yield clear performance improvements, indicating that the benchmark's difficulty cannot be resolved simply by sampling more frames. We also evaluate a subset excluding YouTube-sourced videos and observe similar trends, suggesting that the strong performance of Gemini 3 and GPT 5 is unlikely to be driven primarily by data leakage.

Caption Source	Additional Verifier	F1-Score (%)		
		Self-Verification	Cross-Model Ensemble	Δ (p.p.)
Qwen3-VL 32B	InternVL3.5 38B	5.3	26.2	**+20.9**
InternVL3.5 38B	Keye-VL 1.5 8B	7.7	24.6	**+16.9**
Gemini 3 Flash	GPT 5 mini	21.9	25.1	**+3.2**

Table 4: **Effect of cross-model ensemble on self-verification.** Cross-model ensemble consistently improves F1-Score over self-verification across caption sources. These gains show that combining judgments from other Video-LLMs can compensate for the verification performance observed when models evaluate their own captions.

5 Conclusion

We introduced VidOmni-Bench, a sentence-wise event verification benchmark for fine-grained video understanding in Video-LLMs. By requiring models to verify dense caption sentences across video complexity and duration, VidOmni-Bench reveals model-specific vulnerabilities and self-preference bias. These findings show that even strong models still struggle with reliable, temporally grounded event verification. We hope VidOmni-Bench serves as a rigorous evaluation standard for more faithful fine-grained video understanding.

Limitations

VidOmni-Bench has several limitations that suggest directions for future work. First, verifying dense video captions inevitably involves some ambiguity, especially for complex temporal events. We mitigate this issue by introducing an Unknown category and using a multi-step labeling protocol, but complete agreement can still be difficult in borderline cases. Second, our analysis reveals self-preference bias in Video-LLMs, suggesting that model-based caption generation and evaluation may benefit from more diverse caption sources and cross-verification protocols. Third, the benchmark construction depends on the generative capability of current Video-LLMs and requires human verification to ensure data quality, which may limit scalability. Despite these limitations, VidOmni-Bench provides a useful diagnostic resource for studying current gaps in fine-grained video understanding.

Ethics Statement

AI assistant usage in writing. ChatGPT (Singh et al., 2026) was used to improve readability during manuscript preparation. The authors reviewed and edited all generated text and take full responsibility for the final content.

Dataset construction. To respect the copyright and terms of service of the original video providers, we do not redistribute raw video files for any of the 500 videos in VidOmni-Bench, regardless of source. Instead, we release only video metadata, the generated dense captions, and the annotations. We note that VidOmni-Bench includes videos from anomaly detection datasets that may depict violent or disturbing events, as well as footage that may contain identifiable individuals. We apply no additional filtering beyond the curation already performed by the original providers, and recommend that users exercise discretion accordingly.

Acknowledgements

This research was supported by the Institute of Information &communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. RS-2019-II190079, Artificial Intelligence Graduate School Program(Korea University), 1%; No. RS-2025-25439490, 40%), the KOCCA grant (RS-2024-00345025, 10%), the National Research Foundation of Korea(NRF) grant funded by the Korea government(MSIT)(No. RS-2025-02263628, 34%) This research was supported by JST PRESTO, Japan, Grant Number JPMJPR2523 (5%). This work was partly achieved through the use of SQUID at D3 Center, The University of Osaka (5%). This work was supported in part by the Physical AI Development Support Program by AWS Japan through the provision of computational resources (5%).

References

Shuai Bai, Yuxuan Cai, Ruizhe Chen, Keqin Chen, Xionghui Chen, Zesen Cheng, Lianghao Deng, Wei Ding, Chang Gao, Chunjiang Ge, Wenbin Ge, Zhifang Guo, Qidong Huang, Jie Huang, Fei Huang, Binyuan Hui, Shutong Jiang, Zhaohai Li, Mingsheng Li, and 45 others. 2025. Qwen3-vl technical report. *Preprint*, arXiv:2511.21631.

Dan Ben Ami, Gabriele Serussi, Kobi Cohen, and Chaim Baskin. 2026. Herbench: A benchmark for

multi-evidence integration in video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4505–4514.

Fabian Caba Heilbron, Victor Escorcia, Bernard Ghanem, and Juan Carlos Niebles. 2015. Activitynet: A large-scale video benchmark for human activity understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, Yao Dou, Jaden Park, Jianfeng Gao, Yong Jae Lee, and Jianwei Yang. 2024. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *Preprint*, arXiv:2410.10818.

Lin Chen, Xilin Wei, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Bin Lin, Zhenyu Tang, Li Yuan, Yu Qiao, Dahua Lin, Feng Zhao, and Jiaqi Wang. 2024a. Sharegpt4video: Improving video understanding and generation with better captions. In *Advances in Neural Information Processing Systems*, volume 37, pages 19472–19495. Curran Associates, Inc.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024b. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 24185–24198.

Dima Damen, Hazel Doughty, Giovanni Maria Farinella, Sanja Fidler, Antonino Furnari, Evangelos Kazakos, Davide Moltisanti, Jonathan Munro, Toby Perrett, Will Price, and Michael Wray. 2018. Scaling egocentric vision: The dataset. In *Computer Vision – ECCV 2018*, pages 753–771, Cham. Springer International Publishing.

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

Bo Feng, Zhengfeng Lai, Shiyu Li, Zizhen Wang, Simon Wang, Ping Huang, and Meng Cao. 2025. Breaking down video llm benchmarks: Knowledge, spatial perception, or true temporal understanding? *Preprint*, arXiv:2505.14321.

Joseph L Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Caifeng Shan, and 2 others. 2025. 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)*, pages 24108–24118.

Hongcheng Gao, Jiashu Qu, Jingyi Tang, Baolong Bi, Yue Liu, Hongyu Chen, Li Liang, Li Su, and Qingming Huang. 2025. Exploring hallucination of large multimodal models in video understanding: Benchmark, analysis and mitigation. *Preprint*, arXiv:2503.19622.

Google. 2025. Gemini 3 developer guide (gemini api): Gemini 3 flash. `https://ai.google.dev/gemini-api/docs/gemini-3`. Accessed 2026-08-24.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kristen Grauman, Andrew Westbury, Eugene Byrne, Zachary Chavis, Antonino Furnari, Rohit Girdhar, Jackson Hamburger, Hao Jiang, Miao Liu, Xingyu Liu, Miguel Martin, Tushar Nagarajan, Ilija Radosavovic, Santhosh Kumar Ramakrishnan, Fiona Ryan, Jayant Sharma, Michael Wray, Mengmeng Xu, Eric Zhongcong Xu, and 66 others. 2022. Ego4d: Around the world in 3,000 hours of egocentric video. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18995–19012.

Hyeonjeong Ha, Jinjin Ge, Bo Feng, Kaixin Ma, and Gargi Chakraborty. 2026. Narrativetrack: Evaluating entity-centric reasoning for narrative understanding. In *Computer Vision – ECCV 2026*, pages 175–193, Cham. Springer Nature Switzerland.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. 2020. spacy: Industrial-strength natural language processing in python.

Brian C Hosler, Xinwei Zhao, Owen Mayer, Chen Chen, James A Shackleford, and Matthew C Stamm. 2019. The video authentication and camera identification database: A new database for video forensics. *IEEE access*, 7:76937–76948.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Xiang Yue, Bo Li, Yuanhan Zhang, and Ziwei Liu. 2026. Video-MMMU: Evaluating knowledge acquisition from multidisciplinary professional videos. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 27798–27828, San Diego, California, United States. Association for Computational Linguistics.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Seungone Kim, Juyoung Suk, Shayne Longpre, Bill Yuchen Lin, Jamin Shin, Sean Welleck, Graham Neubig, Moontae Lee, Kyungjae Lee, and Minjoon Seo. 2024. Prometheus 2: An open source language model specialized in evaluating other language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4334–4353, Miami, Florida, USA. Association for Computational Linguistics.

Fanheng Kong, Jingyuan Zhang, Hongzhi Zhang, Shi Feng, Daling Wang, Linhao Yu, Xingguang Ji, Yu Tian, Victoria W., and Fuzheng Zhang. 2025. TUNA: Comprehensive fine-grained temporal understanding evaluation on dense dynamic videos. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1810–1839, Vienna, Austria. Association for Computational Linguistics.

Kyuho Lee, Euntae Kim, Jinwoo Choi, and Buru Chang. 2025. Noah: Benchmarking narrative prior driven hallucination and omission in video large language models. *Preprint*, arXiv:2511.06475.

Baiqi Li, Kangyi Zhao, Ce Zhang, Chancharik Mitra, Jean de Dieu Nyandwi, and Gedas Bertasius. 2026. Timeblind: A spatio-temporal compositionality benchmark for video llms. *Preprint*, arXiv:2602.00288.

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

Yanwei Li, Chengyao Wang, and Jiaya Jia. 2025. Llama-vid: An image is worth 2 tokens in large language models. In *Computer Vision – ECCV 2024*, pages 323–340, Cham. Springer Nature Switzerland.

Zhengqi Li, Tali Dekel, Forrester Cole, Richard Tucker, Noah Snavely, Ce Liu, and William T. Freeman. 2021. Mannequinchallenge: Learning the depths of moving people by watching frozen people. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(12):4229–4241.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, volume 36, pages 34892–34916. Curran Associates, Inc.

Zhihang Liu, Chen-Wei Xie, Bin Wen, Feiwu Yu, JixuanChen, Pandeng Li, Boqiang Zhang, Nianzu Yang, YingluLi, Zuan Gao, Yun Zheng, and Hongtao Xie. 2025. Capability: A comprehensive visual caption benchmark for evaluating both correctness and thoroughness. In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc.

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

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

Kuniaki Saito, Risa Shinoda, Shohei Tanaka, Tosho Hirasawa, Fumio Okura, and Yoshitaka Ushiku. 2026. Haldec-bench: Benchmarking hallucination detector in image captioning. *Preprint*, arXiv:2511.20515.

Fadime Sener, Dibyadip Chatterjee, Daniel Shelepov, Kun He, Dipika Singhania, Robert Wang, and Angela Yao. 2022. Assembly101: A large-scale multi-view video dataset for understanding procedural activities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 21096–21106.

Aaditya Singh, Adam Fry, Adam Perelman, Adam Tart, Adi Ganesh, Ahmed El-Kishky, Aidan McLaughlin, Aiden Low, AJ Ostrow, Akhila Ananthram, Akshay Nathan, Alan Luo, Alec Helyar, Aleksander Madry, Aleksandr Efremov, Aleksandra Spyra, Alex Baker-Whitcomb, Alex Beutel, Alex Karpenko, and 467 others. 2026. Openai gpt-5 system card. *Preprint*, arXiv:2601.03267.

Waqas Sultani, Chen Chen, and Mubarak Shah. 2018. Real-world anomaly detection in surveillance videos. In *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6479–6488.

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

Changli Tang, Yixuan Li, Yudong Yang, Jimin Zhuang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. 2025. video-salmonn 2: Caption-enhanced audio-visual large language models. *arXiv preprint arXiv:2506.15220*.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, Ali Vosoughi, Chao Huang, Zeliang Zhang, Pinxin Liu, Mingqian Feng, Feng Zheng, Jianguo Zhang, Ping Luo, Jiebo Luo, and Chenliang Xu. 2026. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*, 36(2):1355–1376.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2025. Label Studio: Data labeling software. Open source software available from https://github.com/HumanSignal/label-studio.

Jiahao Wang, Yufeng Yuan, Rujie Zheng, Youtian Lin, Jian Gao, Lin-Zhuo Chen, Yajie Bao, Chang Zeng, Yanxi Zhou, Xiao-Xiao Long, Hao Zhu, Zhaoxiang Zhang, Xun Cao, and Yao Yao. 2026. Spatialvid: A large-scale video dataset with spatial annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 42592–42603.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, and Jie Tang. 2025a. Lvbench: An extreme long video understanding benchmark. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 22958–22967.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, Zhaokai Wang, Zhe Chen, Hongjie Zhang, Ganlin Yang, Haomin Wang, Qi Wei, Jinhui Yin, Wenhao Li, Erfei Cui, and 56 others. 2025b. Internvl3.5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *Preprint*, arXiv:2508.18265.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Peng Wu, Jing Liu, Yujia Shi, Yujia Sun, Fangtao Shao, Zhaoyang Wu, and Zhiwei Yang. 2020. Not only look, but also listen: Learning multimodal violence detection under weak supervision. In *Computer Vision – ECCV 2020*, pages 322–339, Cham. Springer International Publishing.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, Fan Yang, Guorui Zhou, Guowang Zhang, Han Shen, Hao Peng, Haojie Ding, Hao Wang, Haonan Fan, Hengrui Ju, and 42 others. 2025b. Kwai keye-vl 1.5 technical report. *Preprint*, arXiv:2509.01563.

Appendix

A Attribution of Icons

The icons used in the figures are sourced from Flaticon (https://www.flaticon.com/). We gratefully acknowledge the creators for providing these visual assets, which are utilized in accordance with their licensing terms to enhance the clarity of our technical illustrations.

B Construction of VidOmni-Bench

B.1 Video Collection

More details of video collection. Before collecting videos for VidOmni-Bench, we first define a taxonomy that categorizes videos according to the complexity of their spatio-temporal dynamics. The taxonomy consists of five complexity types, where each type progressively introduces an additional source of difficulty: motion, fine-grained action, multi-agent binding, and scene transitions.

This design results in five complexity types. Static World videos contain little or no temporal change, so they can largely be understood from static visual content. Body Motion videos introduce temporal changes caused by human or object motion, such as walking, running, or turning. Hand-Object Interaction videos further require understanding fine-grained local actions, such as grasping, placing, or manipulating an object. Crowd Action videos involve multiple agents acting simultaneously, where the model must correctly associate each action with the corresponding subject or object. Finally, Multi-Scene Dynamics videos contain shot changes or multiple scenes, requiring the model to connect events across temporally discontinuous segments.

We note that these types are not intended to be mutually exclusive definitions of intrinsic video complexity. Instead, our taxonomy is designed as a controlled diagnostic dimension, where each type emphasizes a different source of spatio-temporal difficulty. As shown in Figure 16, this taxonomy provides a useful axis for examining how Video-LLMs behave under different video dynamics. Tables 5 and 6 summarize the complexity taxonomy and the public video sources used for each type.

Complexity	Motion	Fine-grained Action	Multi-Agent Binding	Scene Transitions
Type 1	✗	✗	✗	✗
Type 2	✓	✗	✗	✗
Type 3	✓	✓	✗	✗
Type 4	✓	✓	✓	✗
Type 5	✓	✓	✓	✓

Table 5: **Complexity taxonomy of VidOmni-Bench.**

Complexity	Name	Video Source (Public Dataset)
Type 1	Static World	Video-ACID (Hosler et al., 2019) SpatialVID (Wang et al., 2026) Mannequin Challenge (Li et al., 2021)
Type 2	Body Motion	ActivityNet (Caba Heilbron et al., 2015)
Type 3	Hand-Object Interaction	Assembly101 (Sener et al., 2022) Epic-Kitchens (Damen et al., 2018) Ego4D (Grauman et al., 2022)
Type 4	Crowd Action	UCF-Crime (Sultani et al., 2018) Sports-1M (Karpathy et al., 2014)
Type 5	Multi-Scene Dynamics	Video-MME (Fu et al., 2025) XD-Violence (Wu et al., 2020)

Table 6: **Data sources of each complexity type in VidOmni-Bench**. Videos are primarily collected from public datasets. We further supplement underrepresented types and duration ranges with videos sourced from YouTube.

B.2 Dense Video Caption Generation

Captioning setups. Table 7 summarizes the Video-LLMs used for caption generation. Open-source models are deployed with official implementations and publicly available weights. Input Frames Per Second (FPS) and the maximum number of input frames are set following reported configurations from Video-MME (Fu et al., 2025), Video-MMMU (Hu et al., 2026), and LVBench (Wang et al., 2025a) whenever available. Otherwise, we carefully set configurations that cover as much video content as possible within each model's context length. We fix the generation seed for reproducibility. For the Video-LLMs used for dense captioning, we randomly select one of the instructions in Figure 13.

Model	Provider	Open/Closed	FPS	Max Frames	Release
Gemini 3 Flash	Google	Closed	1	-	2026/02
Qwen3-VL 32B	Alibaba	Open	2	2048	2026/01
InternVL 3.5 38B	OpenGVLab	Open	1	32	2025/12
video-SALMONN 2+ 72B	ByteDance	Open	1	768	2025/11
Keye-VL 1.5 8B	Kwai	Open	1	1024	2025/03

Table 7: **Details of Video-LLMs chosen as captioning engines.** We use five Video-LLMs with diverse providers, accessibility, FPS, and maximum number of frames (Max Frames) to generate dense captions.

Instruction given to Video-LLMs in captioning

1. Describe this video in detail, focusing on the spatio-temporal dynamics. Describe exactly how objects and agents move, change, and occupy space over time within the scene.
2. Give a detailed account of everything shown in the video, capturing all visible specifics. Describe events in the exact order they appear over time. Ensure that you describe the sequence of events exactly as they occur, without skipping any steps.
3. Describe the video in detail, paying special attention to how objects and people interact with each other. Capture the precise timing and nature of every contact, movement, and reaction shown in the footage.
4. Thoroughly describe the video's visual narrative, capturing every visible detail from start to end. Emphasize how actions unfold and how the scene transitions logically over time.
5. Provide a thorough description of every detail, explicitly prioritizing spatio-temporal dynamics. Capture all visual elements, and focus on their spatial positions, movement trajectories, and how the scene evolves over time.

Figure 13: **Prompt templates for dense video captioning.** We use five prompt templates to generate dense captions for each video, reducing the potential instruction bias that may arise from relying on a single captioning prompt.

B.3 Sentence-wise Annotation

Annotator. We recruit five annotators specifically for our multi-stage labeling task, which requires meticulous reading and temporal reasoning to identify hallucinated sentences within dense captions. All candidates are required to demonstrate high-level English proficiency and the ability to detect subtle spatio-temporal inconsistencies. These qualifications are rigorously verified through a pilot annotation task involving 10 videos prior to the main annotation phase.

Voting and quality control. The authors review all pilot annotations, and only those who demonstrated high accuracy in identifying hallucinated sentences are designated as trusted annotators for the main task. Each trusted annotator is assigned between 500 and 1,000 video-caption pairs. To ensure objective and robust annotation, each pair is independently evaluated by at least two annotators, including a minimum of one trusted annotator. Once at least two annotations are collected for a sentence, the final label is determined through majority voting. Furthermore, the authors conduct systematic audits for every batch of 100 video-caption pairs. If a batch fails to meet our criteria, annotators are required to re-evaluate the entire set before proceeding.

Inter-annotator agreement. We evaluate inter-annotator agreement on VidOmni-Bench using Fleiss' kappa (Fleiss, 1971) with labels (Correct, Incorrect, Unknown). The overall agreement reaches κ = 0.50 with an observed agreement of P_{bar} = 0.764. While these values indicate a moderate level of agreement, we prioritize data integrity through a rigorous manual adjudication process. To ensure the absolute reliability of the benchmark, authors meticulously review and resolve all disputed labels and cases with low confidence, establishing a high-quality ground truth for subsequent evaluation.

Annotation interface. We customize the Label Studio (Tkachenko et al., 2020-2025) interface for our two-stage annotation pipeline. Figure 17 shows the interface for sentence-wise correctness annotation (Stage 1), while Figure 18 shows the interface for word-level error category annotation (Stage 2).

Word-level error categories. Table 8 provides detailed descriptions for each error categories used during our annotation process. For every sentence labeled as Incorrect, annotators are instructed to pinpoint the specific errors at the word level and classify them according to these categories.

B.4 Edge Case Handling Guidelines

Due to the length and complexity of generated captions, we establish internal guidelines to maintain consistency, particularly for ambiguous or repetitive sentences that are difficult to classify using our error categories. These guidelines function as decision rules to help annotators navigate edge cases systematically.

Handling repetitive sentences. A common failure mode in Keye-VL 1.5 (Yang et al., 2025b) and video-SALMONN 2+ (Tang et al., 2025) is that a model repeatedly generates the same or similar sentences. In such cases, we evaluate the first occurrence based on its correctness. If the video progresses to a new event but the model continues to repeat the previous sentence, all subsequent repetitions are labeled Incorrect and assigned the Frequency error. This ensures the annotation reflects the model's inability to track temporal progression.

Type	Description	Example
Object	Misdescriptions of the object itself including its attributes (color, text, count, etc.).	There are two blue cars, but described as "three red cars."
Relation	Misdescriptions of spatial relationships between objects.	A cup is under the table, but described as "on the table."
Location	Misdescriptions of an object's position.	A cup is on the right side, but described as "on the left."
Direction	Misdescriptions of orientation or direction of an object.	A person is looking to the right, but described as "looking to the left."
Illusion	Mentioning non-existent objects, events, or actions.	No bird is present, but described as "a bird flying in the sky."
Action	Incorrect or incomplete description of actions over time.	A person is walking, but described as "running."
Order	Misdescriptions of the temporal order of events or actions.	Sat down after opening a door, but described as "door opened after sitting down."
Causality	Misdescriptions of causal relationships among events.	Describing that rain was caused by people opening umbrellas.
Frequency	Incorrect tracking of the frequency of repeated actions.	A person jumped once, but described as "jumped three times."
State Change	Misdescribing an object's state transition.	Traffic light changed from red to yellow, but described as "yellow to red."

Table 8: **Word-level error categories in VidOmni-Bench.** We group word-level errors into five static-frame-related categories(Object, Relation, Location, Direction, and Illusion) and five temporal-understanding-related categories(Action, Order, Causality, Frequency, and State Change).

Distinguishing temporal errors. We provide explicit instructions to help annotators distinguish events that are temporally misplaced from those that are entirely hallucinated. For example, if a model describes an event that does not occur at the specified timestamp but happens elsewhere in the video, it is labeled as an Order error rather than an Illusion error. This distinction is crucial for diagnosing whether a model suffers from total hallucination or merely temporal shuffling of frames.

Resolving visual ambiguity. In cases where the visual evidence is too ambiguous to make a definitive judgment, annotators utilize the Unknown category or seek final adjudication through discussions with the authors. This collaborative feedback loop ensures that the final dataset remains a reliable ground truth for fine-grained video understanding.

C Details of Experimental Setups

C.1 Details of Evaluation on VidOmni-Bench

Source of models. For open-weight models, we use checkpoints available on Hugging Face (Wolf et al., 2020) and base our code on the Hugging Face Transformers package. For GPT 5 mini and Gemini 3 Flash, we use their official APIs.

Prompt for sentence-wise event verification. To ensure a fair and consistent evaluation, we apply a unified chat template across all evaluated Video-LLMs. The full prompt template is shown in Figure 14. For each inference, the prompt remains identical except for the specific target caption provided for sentence-wise verification.

Computation. At most four H200 GPUs are used for inference of a single model.

Evaluation configs. As in our captioning pipeline, we set FPS and the maximum number of input frames following each benchmark's reported settings whenever available, and otherwise sample as many frames as possible up to each model's native context limit, as detailed in Table 9.

Model	Provider	Open/Closed	FPS	Max Frames	Release
Gemini 3 Flash	Google	Closed	1	-	2026/02
GPT 5 mini	OpenAI	Closed	-	256	2026/03
Qwen3-VL 32B	Alibaba	Open	2	2048	2026/01
InternVL 3.5 38B	OpenGVLab	Open	1	32	2025/12
Video-SALMONN 2+ 72B	ByteDance	Open	1	768	2025/11
Keye-VL 1.5 8B	Kwai	Open	1	1024	2025/03

Table 9: **Details of Video-LLMs evaluated on VidOmni-Bench.** We list the provider, accessibility, FPS, and maximum number of input frames (Max Frames) for each Video-LLM evaluated on VidOmni-Bench.

D Additional Analysis

D.1 Caption Analysis in VidOmni-Bench

Details for Certificate Coverage. We follow Video-MME (Fu et al., 2025), which defines Certificate Length as the minimum video span required to solve a task. For sentence-wise event verification, we annotate the minimum video intervals needed

Instruction given to Video-LLMs in sentence-wise verification

You are an expert in Video Hallucination Detection. You are given a video and a caption that consists of multiple sentences. Your task is to evaluate each sentence individually to determine if it contains a hallucination based on the visual evidence.

For each sentence, assign a binary label (0 or 1) according to the following definition:

- Label 1: The sentence contains a hallucination. This means it describes something not present or incorrect regarding the video.
- Label 0: The sentence is faithful to the video. This means it correctly describes the visual content.

Please first output the dictionary of labels before explaining the reason for your judgment.

Output format requirements:

1) First line must be a valid JSON dictionary with keys sentence_1, sentence_2, ... and values 0 or 1.

2) Then explain the reason for each sentence.

Caption:

Figure 14: **Sentence-wise event verification instruction.** The prompt asks Video-LLMs to assign a binary label to each caption sentence before providing sentence-level reasoning.

to determine whether each sentence is supported by the video. Certificate Coverage is then computed as the ratio of the annotated Certificate Length to the full video duration. Since annotating Certificate Length is highly labor-intensive, the Certificate Coverage reported in Figure 7 and Table 2 is computed on a randomly sampled subset of 30 videos for each benchmark.

Captions of VidOmni-Bench are more complex than other benchmarks. We compare the informativeness of captions in VidOmni-Bench with those from ARGUS (Rawal et al., 2025), NOAH (Lee et al., 2025), and CAPAbility (Liu et al., 2025). We first conduct a lexical and syntactic analysis using spaCy (Honnibal et al., 2020). As shown in Table 10, captions in VidOmni-Bench contain substantially more words, information units, noun chunks, and verbs than captions from prior benchmarks.

In addition to these surface-level statistics, we further assess caption informativeness through a pairwise LLM-as-a-judge evaluation following

Rubrics for informativeness comparison

1. The response is severely lacking in information. It is vague, overly generic, or omits essential details, leaving the query largely unaddressed.
2. The response provides minimal information. It touches on the topic but lacks depth, with significant gaps or important aspects left uncovered.
3. The response provides a moderate amount of information. It covers the core of the query but may lack depth in some areas or omit secondary details that would add value.
4. The response is informative and substantive. It thoroughly addresses the query with relevant details, with only minor opportunities for additional depth or coverage.
5. The response is highly informative and comprehensive. It fully addresses the query with rich, relevant, and well-elaborated detail, leaving little need for follow-up.

Figure 15: **Rubrics for pairwise caption informativeness judgment.** We use Prometheus-2 7B to compare captions from VidOmni-Bench and other benchmarks under a predefined informativeness rubric. Given two captions and the rubric, the judge selects the more informative caption.

Prometheus-2 (Kim et al., 2024). Given the predefined informativeness rubric in Figure 15, the judge compares each VidOmni-Bench caption with a caption from another benchmark and selects the more informative one. VidOmni-Bench captions win in 96.0–99.6% of pairwise comparisons, suggesting that they provide richer and more detailed event descriptions for sentence-wise event verification.

D.2 Additional Experiments on VidOmni-Bench

YouTube leakage analysis. Table 12 compares performance on the full VidOmni-Bench test set and a subset excluding YouTube-sourced videos. F1-Scores remain consistent across both settings, indicating that the results in Table 3 are not substantially driven by YouTube-sourced videos. This suggests limited impact from potential YouTube data leakage or source-specific bias.

VidOmni-Bench remains challenging beyond frame sampling. Table 11 reports type-wise F1-Scores of Qwen3-VL 32B under different FPS from 1 to 10. Overall, F1-Scores show no consistent improvement as FPS increases, indicating that poor performances in Table 3 are not primarily driven by insufficient frame sampling.

Benchmark	Avg. Words	Info Count	Noun Chunks	Verbs	LLM Judge Win Rate (%) (vs. Ours)
ARGUS (Rawal et al., 2025)	477.5	336.4	134.3	63.1	4.0
NOAH (Lee et al., 2025)	48.0	34.2	13.8	8.0	0.6
CAPAbility (Liu et al., 2025)	3.2	2.4	1.1	0.3	0.4
VidOmni-Bench	578.9	400.6	159.3	82.1	–

Table 10: **Comparison of caption informativeness across benchmarks.** VidOmni-Bench achieves substantially longer captions, higher information counts, and more noun chunks and verbs than prior benchmarks. The last column reports the percentage of pairwise LLM-as-a-judge comparisons in which the baseline caption is preferred over the VidOmni-Bench caption in terms of informativeness.

Video-LLM	FPS	Type 1	Type 2	Type 3	Type 4	Type 5
Qwen3-VL 32B	1	27.32	16.21	22.18	37.08	25.22
	2	15.19	14.44	10.25	33.36	21.22
	4	18.38	15.99	31.75	37.71	22.24
	8	18.38	20.09	20.14	29.04	24.63
	10	18.38	23.06	19.74	28.49	24.02

Table 11: **Effect of frame sampling rate on Qwen3-VL 32B performance across complexity types.**

Aggregated performance across all complexity levels and durations. Figure 16 reports the F1-Score of each Video-LLM across the complexity and duration axes of VidOmni-Bench. The results show that Video-LLM performance varies substantially across different combinations of video dynamics and temporal scale. This reveals multidimensional weaknesses, where models do not fail uniformly but instead exhibit different vulnerabilities depending on the required visual reasoning and temporal coverage.

VidOmni-Bench Subset	Video-LLM	F1-Score (%)
Full Test Set	Gemini 3 Flash	41.69
	GPT 5 mini	28.93
	Qwen3-VL 32B	21.74
w/o YouTube Videos	Gemini 3 Flash	42.46
	GPT 5 mini	29.15
	Qwen3-VL 32B	23.21

Table 12: **YouTube data leakage analysis.** The similar F1-Scores across the two settings suggest that the reported results are not driven by leakage from YouTube.

Analysis of audio modality in verification. To investigate the impact of audio information on sentence-wise event verification, we conduct one-sample t-tests within three audio categories defined by the type of content they convey: Speech (meaningful spoken content), Sound Effect (semantically relevant non-speech sounds), and Background (uninformative audio). As shown in Table 13, for Gemini 3 Flash, audio input significantly improves

Video-LLM	Audio Category	F1-Score Gain (%)	t	Significance
Gemini 3 Flash	Speech	+3.3	+3.9	***
	Sound Effect	+4.2	+3.6	***
	Background	−0.8	−0.9	n.s.
video-SALMONN 2+ 7B	Speech	−0.3	−1.7	n.s.
	Sound Effect	−1.5	−2.5	*
	Background	−0.9	−2.7	**
video-SALMONN 2+ 72B	Speech	−2.2	−4.2	***
	Sound Effect	−2.4	−2.6	*
	Background	−1.6	−2.6	**

Table 13: **F1-Score gain when audio is added by audio category and model.** For each model and audio category, we compute the F1-Score gain (with audio - without audio) and test whether this gain is significantly different from zero using a one-sample t-test. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$, n.s.: not significant $p \geq .05$.

F1-Score in both Speech and Sound Effect, while causing no significant change in the Background. In contrast, F1-Scores of video-SALMONN 2+ decline across these categories. These results suggest that while Gemini effectively leverages audio as a complementary signal, video-SALMONN 2+ tends to treat audio as a distractor.

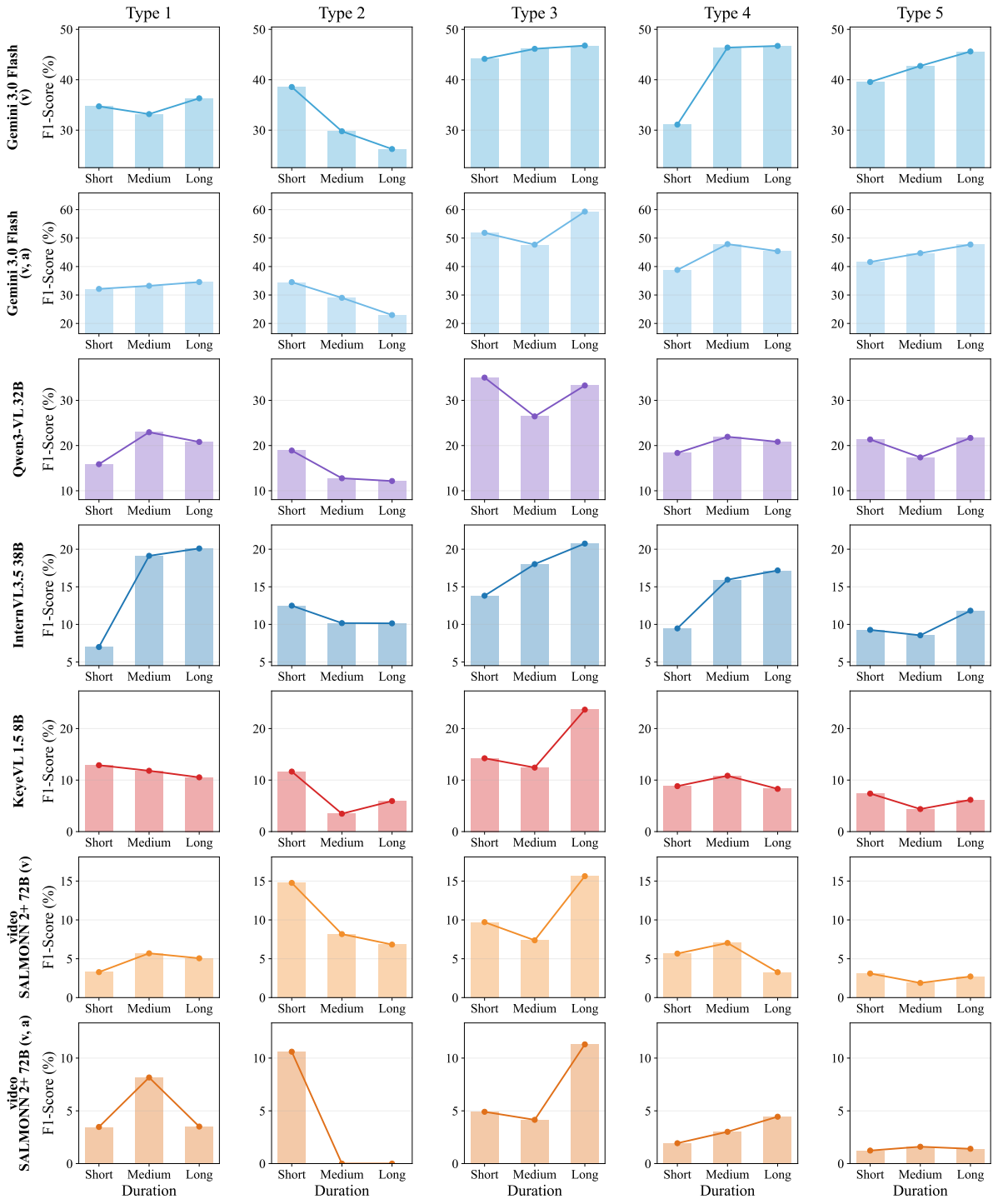

Figure 16: **Performances of Video-LLMs across all complexity types and durations**. The results reveal multidimensional and model-specific weaknesses, indicating that current Video-LLMs exhibit different failure patterns depending on both video dynamics and temporal scale.

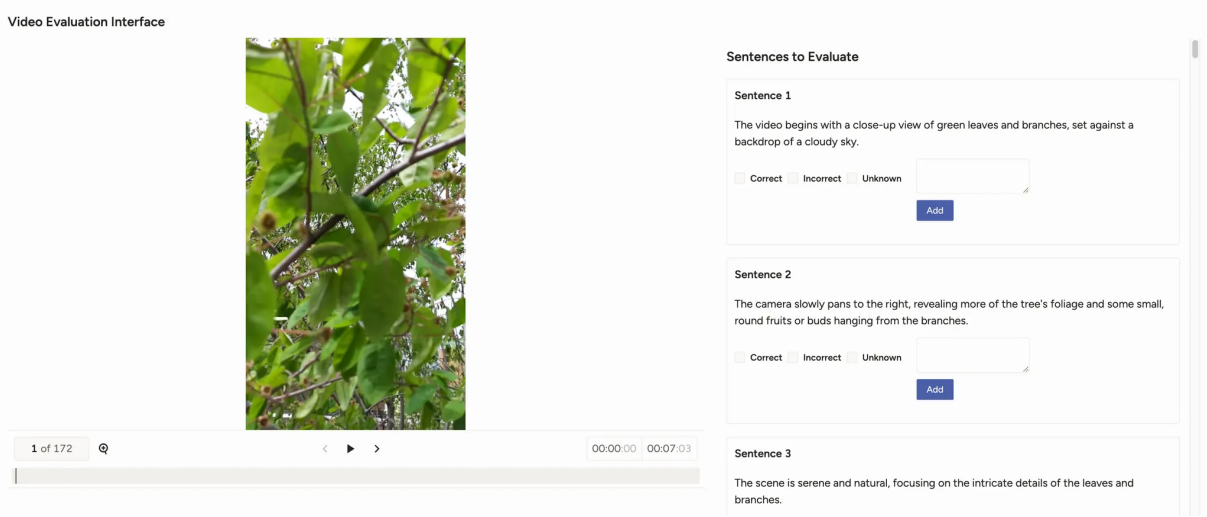

Figure 17: **Annotation interface for sentence-wise correctness (Stage 1)**. Annotators label each caption sentence as Correct, Incorrect, or Unknown based on the video evidence. An additional memo field is provided to collect detailed notes and ambiguous cases during dataset construction.

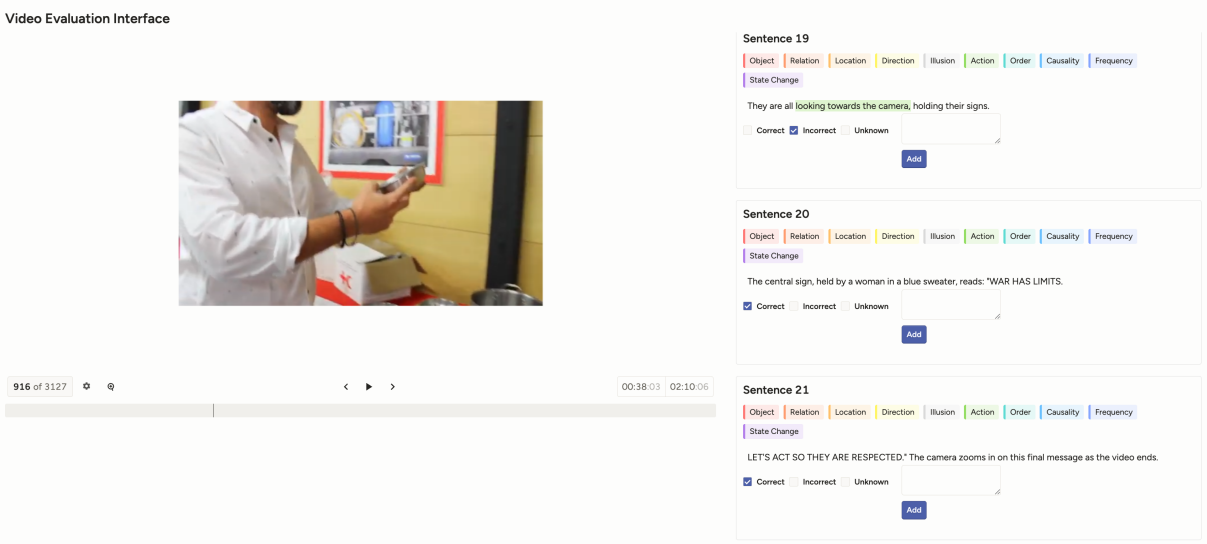

Figure 18: **Annotation interface for word-level error category (Stage 2).** For each sentence labeled as Incorrect in Stage 1, annotators identify the erroneous words or phrases and assign them to predefined error categories, enabling fine-grained analysis of hallucination types.