MOMENT-VIDEO: DIAGNOSING TEMPORAL FIDELITY OF VIDEO MLLMS ON MOMENTARY VISUAL EVENTS

Anonymous authors
Paper under double-blind review

ABSTRACT

Video multimodal large language models (MLLMs) have made rapid progress in general and long-form video understanding, yet their ability to capture and use brief, answer-critical visual evidence remains underexplored. Momentary actions or state transitions may determine an answer while occupying only a small fraction of the video. Performance on broad video understanding tasks alone does not establish whether models can reliably identify and use this temporally localized evidence. We introduce MOMENT-VIDEO, a benchmark for evaluating temporal fidelity through momentary visual event understanding. Questions are constructed around visually observable events or event sequences, requiring models to notice, count, describe, or reason about transient evidence. The benchmark contains 1,000 human-verified video-QA pairs across seven domains and 25 subcategories, covering four tasks: Temporal Occurrence, Temporal Counting, Action Description, and Temporal Reasoning. We evaluate 34 proprietary and open-source MLLMs. Under the main evaluation settings, the best-performing model, Seed-2.0-Pro, achieves 39.6% overall accuracy, while most open-source models remain below 25%. Diagnostic analyses show that denser frame sampling improves some models, but substantial errors persist even when sampled frames cover all annotated necessary event intervals in an audited subset. Performance also declines on longer videos, suggesting additional temporal-localization challenges. By exposing these limitations, MOMENT-VIDEO provides a diagnostic benchmark to guide the evaluation and development of video MLLMs that more reliably understand brief, answer-critical visual events.

1 INTRODUCTION

Video multimodal large language models (MLLMs) have recently made substantial progress in dynamic visual understanding. By extending image-language models to videos through frame sampling, temporal positional encoding, visual-token compression, and multimodal instruction tuning, recent systems achieve strong performance in video question answering, action recognition, temporal reasoning, and long-video comprehension Bai et al. (2025); Zhang et al. (2025a); Wang et al. (2025b); Sun et al. (2025). More recent omni-modal and reasoning-oriented models further enhance video understanding with audio-visual inputs, chain-of-thought reasoning, reinforcement learning, and tool-augmented inference Kimi Team et al. (2025); Qwen Team (2026c); Hong et al. (2025b); Feng et al. (2025); Li et al. (2025); Yan et al. (2025). Despite these advances, the ability to use brief but decisive visual evidence remains insufficiently evaluated.

Many real-world video questions depend on *momentary visual events*: visually observable actions or state transitions over a short duration, sometimes only a few frames. Examples include a pedestrian suddenly entering the road (Li et al., 2026a), a machine exhibiting momentary abnormal motion (Di Paco et al., 2026), a GUI element flashing after a command, or a game character performing a short attack that determines the outcome. Although these events occupy only a brief portion of the video, they are often decisive for the answer. Sparse sampling can omit critical frames, leaving models without direct observations of the queried event. Correct answers may then depend on indirect contextual cues. This motivates *temporal fidelity*: the ability to capture, retain, and reason over short-lived, answer-critical visual evidence. We assess this capability through model answers, without directly probing internal representations.

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.

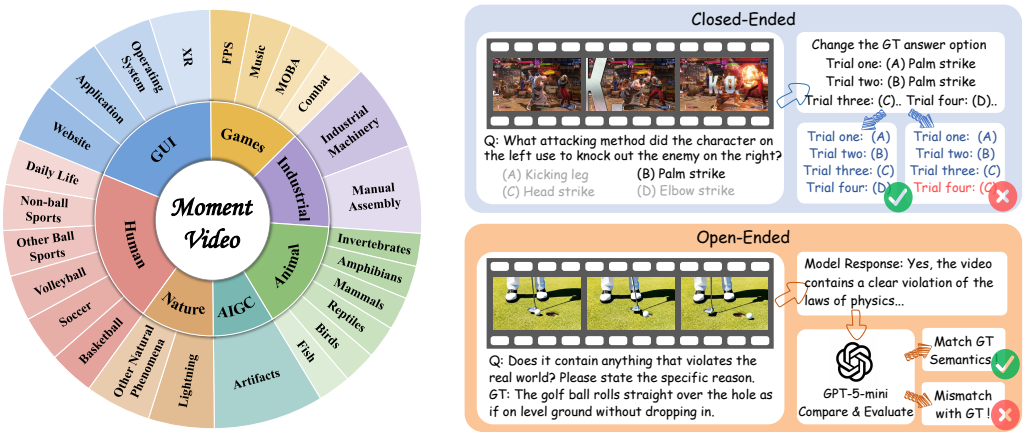


Figure 1: Overview of MOMENT-VIDEO. Left: domain and subcategory taxonomy. Right: evaluation protocols for closed-ended and open-ended questions.

Existing video benchmarks rarely isolate this capability. General-purpose, long-video, and reasoning-oriented benchmarks Li et al. (2024); Fu et al. (2025a; 2026); Wang et al. (2025a); Wu et al. (2024); Zhou et al. (2025); Chandrasegaran et al. (2024); Mangalam et al. (2023); Yeh et al. (2026); Li et al. (2026b); Liu et al. (2026); Hu et al. (2025); Zhao et al. (2025) assess scene understanding, temporal aggregation, or persistent objects. Motion-centric benchmarks Hong et al. (2025a); Tu et al. (2025) evaluate motion perception or description, but their questions generally do not focus on brief, sampling-sensitive events. Thus, it remains unclear whether current MLLMs can reliably detect, count, describe, and reason about transient, momentary visual events.

To address this gap, we introduce MOMENT-VIDEO, a benchmark for diagnosing temporal fidelity via momentary visual event understanding. It contains 1,000 human-verified video-QA pairs across 7 domains and 25 fine-grained subcategories, covering real-world and virtual scenarios including AI-generated content, GUI recordings, natural scenes, industrial environments, games, human activities, and animals. Each question is grounded in a localized momentary event, requiring answers to depend on event evidence rather than global context or external knowledge. A small subset of questions involve transient interface text or observable animal behavior; these cues are included only when they are visually grounded and answer-critical.

Questions in MOMENT-VIDEO are organized into four complementary task types. *Temporal Occurrence (TO)* evaluates whether a model can notice that a brief event or state transition happens. *Temporal Counting (TC)* evaluates whether a model can track and count transient actions, object changes, or repeated event occurrences. *Action Description (AD)* evaluates whether a model can characterize the dynamics of a momentary event, such as its direction, trajectory, target, interaction, or state change. *Temporal Reasoning (TR)* evaluates whether a model can integrate the pre-event state, the momentary event, and the post-event state to infer the event's consequence. Together, these task types form a progressive evaluation of temporal fidelity: from capturing the existence of a brief event, to preserving its dynamic details, to using it for state-change reasoning.

We evaluate 34 proprietary and open-source MLLMs. The best-performing model, Seed-2.0-Pro, achieves 39.6% overall accuracy, while most open-source models remain below 25%. Task-wise, AD achieves the strongest average performance, while TC is consistently the most challenging; performance on TO and TR is more model-dependent and comparable overall. Category-wise results vary across domains; scene complexity, small objects, and fast interactions are possible contributors to these differences.

Diagnostic analyses of frame rate and video duration show that denser sampling improves some models but does not fully resolve the bottleneck. Accuracy is also lower on longer videos, consistent with greater difficulty locating brief evidence. These findings reveal a temporal fidelity limitation in current video MLLMs, motivating more fine-grained, event-aware temporal processing.

Our contributions are summarized as follows:

(1) We introduce MOMENT-VIDEO to evaluate temporal fidelity through brief, answer-critical events, with 1,000 human-verified video-QA pairs across 7 domains and 25 subcategories.

(2) We design four complementary tasks (TO, TC, AD, TR) to assess event detection, counting, description, and reasoning over event-induced state changes.

(3) We evaluate 34 MLLMs and analyze frame rate and video duration. Denser sampling helps some models, yet errors persist even when sampled frames hit all necessary event intervals in an audited subset. Accuracy also declines on longer videos, suggesting additional challenges in locating brief, answer-critical evidence.

2 RELATED WORK

Advancements in Video MLLMs. Recent video MLLMs have evolved from simple frame-based extensions of image-language models to more capable systems with temporal encoding, dynamic visual processing, visual-token compression, and large-scale multimodal instruction tuning (Bai et al., 2025; Zhang et al., 2025a; Wang et al., 2025b; Sun et al., 2025). These models typically transform sampled frames or video clips into temporally ordered visual tokens and align them with LLMs for video question answering, action recognition, temporal reasoning, and long-video comprehension. Meanwhile, omni-modal and reasoning-oriented models further extend video understanding with audio-visual inputs, thinking-style reasoning, reinforcement learning, and tool-augmented inference (Qwen Team, 2026c; Kimi Team et al., 2025; Hong et al., 2025b; Feng et al., 2025; Li et al., 2025; Yan et al., 2025). Despite these advances, reliably capturing and using brief, answer-critical visual events remains insufficiently evaluated, motivating the focus of MOMENT-VIDEO on momentary-event understanding.

Benchmarks for Video MLLMs. A growing number of benchmarks evaluate video MLLMs. MVBench (Li et al., 2024), Video-MME (Fu et al., 2025a), and Video-MME-v2 (Fu et al., 2026) assess broad video understanding across scenarios and tasks. Other benchmarks emphasize long-video comprehension (Wang et al., 2025a; Wu et al., 2024; Zhou et al., 2025; Chandrasegaran et al., 2024), egocentric understanding (Mangalam et al., 2023), multi-view or cross-video reasoning (Yeh et al., 2026; Li et al., 2026b), or complex video reasoning (Liu et al., 2026; Hu et al., 2025; Zhao et al., 2025). MotionBench (Hong et al., 2025a) evaluates fine-grained motion perception, FAVOR-Bench (Tu et al., 2025) includes closed-ended and open-ended motion understanding, and VideoAds (Zhang et al., 2025b) studies fast-paced advertisements. These settings can contain brief events; MOMENT-VIDEO makes dependence on such evidence an explicit construction criterion. Annotators identify the target event or event sequence before writing questions and reject questions answerable without the critical moment. Across natural and virtual domains, four tasks test event detection, counting, description, and consequence inference. The distinction lies in requiring localized evidence to determine the answer, beyond selecting videos with motion or rapid scene changes. Table 1 summarizes the design differences.

3 BENCHMARK DESIGN

MOMENT-VIDEO is designed to evaluate the temporal fidelity of video MLLMs through momentary visual event understanding. The final benchmark contains 1,000 videos paired with 1,000 human-verified question-answer pairs, spanning 7 high-level domains and 25 fine-grained subcategories. We illustrate the domain taxonomy and the evaluation formats used in the benchmark in Fig. 1.

3.1 WHAT MAKES AN EVENT MOMENTARY?

We define a *momentary visual event* as a brief, visually observable action or state transition whose temporally localized evidence is critical to answering the question. Missing this moment can change the answer even when the surrounding scene is understood. Here, momentariness describes temporal brevity and answer criticality, without imposing a fixed duration threshold. We randomly selected 300 samples through stratified sampling and manually annotated the start and end times of all answer-critical events, yielding 609 intervals. Of these, 87.7% lasted at most 1 second, 97.9% at most 2

Table 1: Comparison with existing video benchmarks along dimensions for momentary visual event understanding. "Partial" means the property appears in part but is not the central design target. "Fully Human-curated" means video selection, question writing, and answer annotation are manual.

Benchmark	#Videos	Primary Focus	Momentary-event Focus	Fully Human-curated	Open-ended	Frame-sampling Analysis
MVBench	4,000	General understanding	✗	✗	✗	✗
Video-MME	900	Comprehensive understanding	✗	✓	✗	✗
Video-MME-v2	800	Robust video understanding	✗	✓	✗	✗
LVBench	103	Long-video understanding	✗	Partial	✗	✗
MotionBench	5,385	Motion perception	Partial	Partial	✗	✓
FAVOR-Bench	1,776	Motion description	Partial	Partial	✓	✗
VideoAds	200	Fast-paced advertisements	Partial	Partial	✗	✓
MOMENT-VIDEO	**1,000**	**Momentary visual events**	✓	✓	✓	✓

seconds, and 71.4% occupied at most 10% of the corresponding video duration (Appendix C). To operationalize this definition during dataset construction, we impose the following criteria:

1. Localized evidence. The decisive evidence for answering the question must be concentrated around a specific moment in the video, rather than being distributed across the entire clip. For example, the answer may depend on the instant when a ball changes direction, an object falls, a GUI element appears, or an animal suddenly moves.

2. Transient visual change. The target event must involve a temporary action, interaction, or state transition. Questions that can be answered from a persistent object attribute, a static scene layout, or the overall topic of the video are excluded. This requirement ensures that the benchmark evaluates dynamic visual evidence rather than static image understanding.

3. Sampling sensitivity. The event should have a non-trivial chance of being missed under sparse frame sampling. When input frames are sampled sparsely, a brief event may fall between sampled timestamps and be absent from the visual input.

4. Evaluability. Each question must have a well-defined answer. We use open-ended questions when the reference answer is semantically convergent and can be judged reliably. When the answer is difficult to standardize in free-form language, we use a multiple-choice format with manually designed distractors that are visually plausible and semantically close to the correct answer.

Together, these criteria distinguish MOMENT-VIDEO from benchmarks that evaluate generic motion recognition or general video-level semantics. The key construction constraint is not simply that the video contains motion, but that the answer depends on brief, localized, visually grounded evidence.

3.2 DATASET CONSTRUCTION

We construct MOMENT-VIDEO through a multi-stage pipeline consisting of domain taxonomy design, video collection, event-centric temporal grounding and human-expert annotation (Appendix Fig. 5).

Domain taxonomy. To ensure broad coverage of momentary visual events, we build a hierarchical taxonomy with 7 high-level domains: AIGC, GUI, Nature, Industry, Games, Human, and Animals. These domains cover diverse visual styles, motion patterns, scene layouts, and event types. Each domain is divided into fine-grained subcategories. For example, the Games domain includes FPS games, MOBA games, combat games, and music games; the Human domain includes ball sports, non-ball sports, and daily-life activities; and the Animals domain includes birds, mammals, reptiles, amphibians, fish, and invertebrates. This taxonomy allows us to evaluate whether models can generalize momentary event understanding across natural videos, screen recordings, synthetic videos, and complex interactive environments. The category distribution is shown in Fig. 1, and the keyword cloud in Fig. 2 further illustrates the semantic coverage of MOMENT-VIDEO.

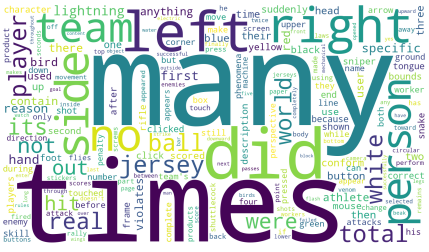


Figure 2: Keyword cloud of frequent semantic concepts in MOMENT-VIDEO.

Video collection. We source candidate videos from public online sources (Zhao et al., 2026; Chen et al., 2024; Hunt, 2020; Build AI, 2025; Li et al., 2021; Safdarnejad et al., 2015; Sun et al., 2020; Soomro et al., 2012; Ng et al., 2022) and self-collected recordings, each required to feature at least one visually clear momentary event. We filter out low-quality videos (e.g., severe blur, low resolution, heavy occlusion) and those relying on audio, subtitles, or external context for comprehension. Furthermore, we remove samples where the answer can be trivially inferred from persistent textual cues or metadata. When necessary, irrelevant visual text is cropped or masked to mitigate OCR-based shortcuts. The final dataset encompasses natural and synthetic domains, including real-world footage, screen recordings, gameplay, and AI-generated videos. Further details are provided in Appendix A.

Event-centric temporal grounding. For each candidate video, annotators first identify the target momentary event before writing the question. This step is essential: the benchmark is constructed from answer-critical events to questions, rather than from generic clips to broad questions. Because many target events are brief and lack reliable frame-level boundaries, we use weak temporal grounding instead of requiring dense start and end timestamps. Annotators record the approximate critical moment, involved objects or agents, the event description, the event count when applicable, and relevant pre-event and post-event states. This information is used for annotation and quality control, but is not provided to evaluated models during testing. When a video contains multiple brief events, annotators specify the queried event or event sequence and the evidence needed to answer the question, including multiple occurrences for counting or reasoning. A sample is rejected or revised if the event cannot be localized, if competing events yield different answers, or if the question is answerable without the key moment.

Human-expert annotation. Based on localized target events, annotators formulate a question and reference answer for each video. Each sample is labeled with a domain, subcategory, task type, and answer format. All samples are manually constructed and cross-checked to ensure clear and robust Q&A pairs. For open-ended questions, the reference answer is a concise semantic description rather than an exhaustive video caption. For choice questions, annotators create visually plausible distractors close to the correct option, avoiding those removable by common sense, option length, or language priors. We further conduct a human solvability study on a benchmark subset to verify that questions are answerable and unambiguous. Further details on dataset statistics, annotation quality control, and human validation are provided in Appendices C, D, and E, respectively.

3.3 Task Taxonomy

We organize questions into four task types: Temporal Occurrence, Temporal Counting, Action Description, and Temporal Reasoning, as shown in Fig. 3. These types decompose temporal fidelity into progressively richer capabilities: noticing brief events, quantifying transient changes, characterizing event dynamics, and reasoning over event-induced state changes. Task formulations and statistics are provided in Appendix C.

Temporal Occurrence (TO). Temporal Occurrence assesses a model's ability to identify momentary events or state transitions. Questions in this category probe the presence, absence, or immediate change of a transient visual event. Typical examples include detecting a falling object, a sudden animal movement, a post-operation GUI update, or an unexpected physical anomaly. This task tests the most basic requirement of temporal fidelity: the model must notice the critical moment.

Temporal Counting (TC). Temporal Counting assesses a model's ability to count transient objects, actions, or event occurrences within a brief visual sequence. Distinct from static object counting, this task necessitates tracking dynamic changes over time while preventing omissions or redundant counts. Typical queries involve quantifying specific actions, state changes, or motions within a critical timeframe. Notably, this task is highly sensitive to sparse sampling, since the omission of even a single brief event can fundamentally alter the final prediction.

Action Description (AD). Action Description evaluates whether a model can characterize the dynamics of a momentary event. Rather than only determining that an event occurs, the model must describe how it unfolds, including its direction, trajectory, target, interaction, motion pattern, or state transition. For example, a question may ask which direction an object moves after impact, which player performs an action, what interaction happens between two agents, or what interface state changes after a brief operation. This task tests whether the model can accurately report the event-specific dynamic details needed to answer the question.

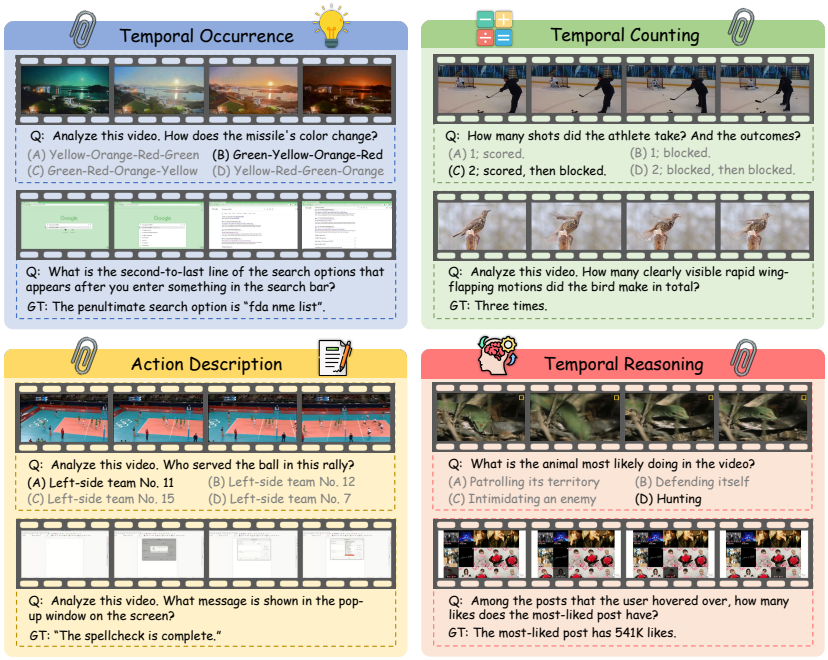


Figure 3: Examples of the four task types in MOMENT-VIDEO. For each type, we show two questions: one closed-ended and one open-ended.

Temporal Reasoning (TR). Temporal Reasoning evaluates whether a model can integrate the pre-event state, momentary event, and post-event state to infer a consequence. Questions in this category require more than event recognition. The model must compare states before and after the event, track how the event changes the scene, and infer the resulting condition. For example, after observing several objects initially present and some quickly leaving, falling, or changing state, the model may need to infer how many remain, what state the scene reaches, or why a subsequent change occurs.

3.4 EVALUATION PROTOCOL AND SCORING

MOMENT-VIDEO contains both multiple-choice and open-ended questions. Each question is converted into a binary correctness label according to its answer format. The overall benchmark score is computed as the average accuracy over all 1,000 samples.

Shuffle-robust multiple-choice evaluation. All multiple-choice questions have exactly one correct option. Each is evaluated under as many shuffled option orders as it has options. A prediction is counted as correct only when the model consistently selects the semantically correct option across all shuffled orders. Formally, for a multiple-choice sample i with ground-truth answer $a_i^\star$, let $\hat{a}_{i,s}$ denote the model prediction under the s-th shuffled option order. The sample is scored as correct only if

$$\mathrm{Acc}_{\mathrm{MC}}(i) = \mathbf{1}\left[\bigwedge_{s=1}^{S_i} \hat{a}_{i,s} = a_i^\star\right], \tag{1}$$

where S_i equals the number of options in sample i (e.g., four evaluations for four options). This shuffle-robust criterion reduces option-position bias and discourages models from relying on superficial answer ordering. When a model outputs the option content instead of the option letter, we map the response to the corresponding option by semantic matching.

Open-ended semantic evaluation. For open-ended questions, exact string matching is insufficient because semantically equivalent answers may be expressed in different forms. We therefore use an LLM-as-judge protocol to evaluate semantic consistency between the model prediction and the reference answer. The evaluator receives the reference answer, and the model prediction, and outputs a binary correctness label. It is instructed to judge only answer equivalence and not to introduce external knowledge or reinterpret the video content. Vague, incomplete, contradictory, or overly

broad answers are marked as incorrect. The full judge prompt and the human validation of judge reliability are provided in Appendix I and F.

4 EXPERIMENTS

4.1 EVALUATION SETUP

We summarize the evaluated models and video input settings below; Appendix B details the evaluation protocols.

Models: We evaluate both proprietary and open-source video MLLMs on MOMENT-VIDEO to examine their ability to capture, preserve, and reason over brief, answer-critical visual evidence. Proprietary models: Seed-2.0 variants (ByteDance Seed Team, 2026), Gemini series (Google DeepMind, 2025), GPT-5.4 (OpenAI, 2026). Open-source models: Qwen series (Qwen Team, 2026d; Bai et al., 2025; Qwen Team, 2026a;b;c), DeepSeek-v4.1-Flash (Xu et al., 2026), InternVL3.5 (Wang et al., 2025b), Kimi 2.6 (Moonshot AI, 2026), and MIMO-v2.5 (XiaomiMiMo, 2026), LLaVA-Video (Zhang et al., 2024), VideoLLaMA3 (Zhang et al., 2025a), GLM (Hong et al., 2025b), Gemma (Gemma Team, Google DeepMind, 2026), Keye-VL (Yang et al., 2025), and VITA (Fu et al., 2025b) models.

Video input processing: For video-native models, we directly feed the original video clip into the model whenever its inference interface supports video input. For models that do not support video input, we convert each video into a sequence of sampled frames and provide them as multi-image input. Most image-based models are evaluated with 1 FPS sampling. If the sampled frames exceed the specified frame budget, we uniformly subsample them across the full video duration.

4.2 MAIN EXPERIMENTAL RESULTS

Overall Performance.

Table 2 summarizes the main results on MOMENT-VIDEO. Overall, current MLLMs show substantial difficulty in momentary visual event understanding. Among proprietary models, Seed-2.0-Pro achieves the best overall accuracy of 39.6%, followed by Seed-2.0-Lite with 31.3% and Seed-2.0-Mini with 27.8%. Other proprietary models, including Gemini-series models and GPT-5.4, obtain accuracies between 20.4% and 26.9%. These results indicate that stronger video-capable models capture a meaningful portion of momentary visual events, but performance remains far from saturated. Appendix G.1 details 8 FPS results and model-specific frame budgets.

For open-source models, Kimi-2.6 achieves the strongest overall performance with 24.9%, followed by Qwen3.5-27B with 24.1%, Qwen3.5-397B-A17B with 22.8%, DeepSeek-v4.1-Flash with 22.6%, and Qwen3.6-27B with 22.5%. MIMO-v2.5 obtains 22.1%, while several Qwen variants obtain around 20%–21%. Many video-oriented open-source models, including Qwen3-VL, InternVL3.5, LLaVA-Video, VideoLLaMA3, and VITA variants, remain below 18%. This suggests that model scale or accepting video input alone does not guarantee reliable momentary visual event understanding. The clear gap between the best proprietary and open-source models further indicates that current open-source MLLMs still have considerable room for improvement on temporally localized visual evidence.

Task-wise Observations. Across task types, Action Description generally achieves the highest average performance, whereas Temporal Counting is consistently the most challenging; results on Temporal Occurrence and Temporal Reasoning vary more across models. For example, Seed-2.0-Pro achieves 50.37% on TO and 47.08% on AD, but only 31.14% on TC. Similar trends appear in open-source models: Kimi-2.6 reaches 36.69% on AD and 30.59% on TR, but 15.89% on TC, while Qwen3.5-27B obtains 32.47% on AD and 19.07% on TC. These results suggest that models are relatively effective at detecting and describing brief events, but struggle to count repeated transient actions. Temporal Reasoning also remains difficult; its questions require relating states before, during, and after a brief event.

Category-wise Observations. Performance also varies substantially across visual domains. Models generally perform better on Nature and Animal videos: Seed-2.0-Pro reaches 69.44% on Nature and 55.00% on Animal, while Gemini-3.1-Flash-Lite achieves the highest proprietary-model Animal accuracy of 57.00%. Among open-source models, MIMO-v2.5 obtains the best Nature accuracy of

Table 2: Main results on **MOMENT-VIDEO**. The **bold** and underlined numbers indicate the best and second-best performance within each group. "Default" means that the model determines which frames to sample and how many frames to use by itself.

Model	Input Settings	Task Type(%)				Video Category(%)							Overall(%)
		TO	TC	AD	TR	AIGC	GUI	Nature	Industry	Games	Human	Animal	
Proprietary Models													
Seed-2.0-Pro	default	**50.37**	**31.14**	**47.08**	**42.35**	**45.57**	**33.94**	**69.44**	**41.32**	**39.37**	**27.20**	55.00	**39.6**
Seed-2.0-Lite	default	34.81	25.64	36.04	40.00	22.78	27.98	51.39	30.58	25.62	26.80	52.00	31.3
Seed-2.0-Mini	default	32.59	22.88	33.77	25.88	24.05	22.94	58.33	29.75	22.50	20.80	43.00	27.8
Gemini-3.1-Pro	1fps	32.59	19.70	33.44	34.12	27.85	27.52	44.44	18.18	24.37	19.20	46.00	26.9
Gemini-3-Flash	1fps	41.48	18.43	31.82	32.94	**45.57**	20.64	43.06	15.70	20.62	23.20	47.00	26.9
Gemini-3.1-Flash-Lite	1fps	34.07	18.64	30.52	27.06	27.85	22.02	41.67	22.31	20.00	14.00	**57.00**	25.1
GPT-5.4	1fps	21.48	12.08	30.19	29.41	6.33	23.85	40.28	4.96	18.75	18.00	37.00	20.4
Open-source Models													
Kimi-2.6	1fps	25.93	15.89	**36.69**	30.59	15.19	**37.16**	41.67	12.40	16.25	**20.40**	34.00	**24.9**
Qwen3.5-27B	1fps	20.74	**19.07**	32.47	27.06	11.39	21.56	41.67	18.18	18.75	20.00	**53.00**	24.1
Qwen3.5-397B-A17B	1fps	27.41	14.83	32.79	23.53	16.46	23.39	47.22	18.18	17.50	14.80	43.00	22.8
DeepSeek-v4.1-Flash	1fps	**30.37**	14.62	28.90	**31.76**	**27.85**	32.57	41.67	9.92	12.50	17.20	28.00	22.6
Qwen3.6-27B	1fps	19.26	17.16	31.82	23.53	10.13	26.61	41.67	19.01	16.25	14.40	44.00	22.5
MIMO-v2.5	default	18.52	17.16	30.84	23.53	2.53	24.31	**52.78**	19.01	**19.38**	14.80	37.00	22.1
Qwen3.5-122B-A10B	1fps	25.19	14.19	27.60	23.53	18.99	22.02	40.28	17.36	13.75	14.40	35.00	20.6
Qwen3.6-35B-A3B	1fps	17.04	14.41	29.22	24.71	3.80	18.35	41.67	**22.31**	15.62	15.20	39.00	20.2
Qwen3.5-35B-A3B	1fps	20.00	16.10	26.30	18.82	8.86	19.27	45.83	19.83	13.75	12.80	40.00	20.0
Gemma-4-31B	1fps	**30.37**	13.56	24.35	22.35	**27.85**	11.93	41.67	17.36	15.62	15.60	36.00	19.9
Qwen3.5-9B	1fps	18.52	13.35	26.62	18.82	10.13	17.43	40.28	12.40	11.87	14.80	40.00	18.6
Qwen3.5-4B	1fps	19.26	13.77	21.43	17.65	8.86	16.97	38.89	13.22	15.00	11.20	32.00	17.2
Qwen3-VL-235B-A22B	1fps	15.56	12.08	23.38	23.53	5.06	16.97	38.89	10.74	10.62	14.00	36.00	17.0
InternVL3.5-30B-A3B	1fps	12.59	16.10	21.75	11.76	3.80	10.55	38.89	15.70	15.00	18.40	27.00	17.0
Qwen3-VL-30B-A3B	1fps	18.52	13.14	20.45	20.00	8.86	14.22	43.06	14.05	15.62	9.60	32.00	16.7
InternVL3.5-8B	1fps	15.56	15.04	20.13	15.29	3.80	11.93	38.89	13.22	13.75	14.00	37.00	16.7
LLaVA-Video-72B	1fps	13.33	15.04	18.83	17.65	0.00	8.26	34.72	19.83	16.25	15.20	31.00	16.2
InternVL3.5-241B-A28B	1fps	19.26	11.02	20.45	18.82	1.27	12.39	40.28	9.09	11.87	13.60	36.00	15.7
Keye-VL-1.5-8B	1fps	14.81	12.92	20.45	14.12	0.00	10.09	29.17	17.36	14.37	14.40	33.00	15.6
InternVL3.5-4B	1fps	14.81	13.14	17.86	7.06	3.80	10.55	29.17	14.05	13.75	10.80	30.00	14.3
GLM-4.6V	1fps	15.56	8.69	19.81	21.18	5.06	10.55	31.94	7.44	11.25	12.40	33.00	14.1
Qwen3-VL-8B	1fps	15.56	10.59	17.86	16.47	8.86	13.30	41.67	9.92	6.88	8.80	29.00	14.0
Qwen3-VL-4B	1fps	14.81	8.69	19.48	21.18	1.27	14.22	41.67	7.44	9.38	9.20	30.00	13.9
LLaVA-Video-7B	1fps	11.11	13.14	13.96	7.06	0.00	8.72	23.61	8.26	15.62	12.80	23.00	12.6
VideoLLaMA3-7B	1fps	7.52	13.69	11.72	8.33	0.00	10.44	25.35	4.96	10.14	15.64	15.00	11.8
GLM-4.6V-Flash	1fps	11.11	7.20	14.29	20.00	1.27	8.72	43.06	13.22	10.00	6.00	12.00	11.0
VITA-1.5	1fps	19.26	8.49	13.42	10.59	0.00	5.31	25.00	8.26	11.25	8.80	26.00	10.6

Table 3: Frame-rate ablation on **MOMENT-VIDEO** at 1/5/8/16 FPS. Frame-sequence inputs use at most 64 frames at 1 FPS and 128 at 5/8/16 FPS. Bold marks each model's best result per metric.

Model	Frame Rate	TO(%)	TC(%)	AD(%)	TR(%)	Overall(%)
Gemini-3.1-Pro Google DeepMind (2025)	1 FPS	32.59	19.70	33.44	34.12	26.9
	5 FPS	**42.22**	**30.08**	**48.05**	42.35	**38.3**
	8 FPS	40.00	27.75	45.45	44.71	36.3
	16 FPS	40.00	27.97	45.78	**45.88**	36.6
Gemini-3-Flash Google DeepMind (2025)	1 FPS	**41.48**	18.43	31.82	32.94	26.9
	5 FPS	34.07	27.75	**49.35**	**38.82**	36.2
	8 FPS	40.00	**27.97**	49.03	**38.82**	**37.0**
	16 FPS	39.26	25.42	46.43	**38.82**	34.9
Qwen3.5-27B Qwen Team (2026d)	1 FPS	20.74	19.07	32.47	**27.06**	24.1
	5 FPS	17.78	18.64	36.69	**27.06**	24.8
	8 FPS	22.22	**19.92**	36.36	20.00	25.3
	16 FPS	**24.44**	18.01	**38.64**	23.53	**25.7**
Qwen3.5-397B-A17B Qwen Team (2026d)	1 FPS	27.41	14.83	32.79	23.53	22.8
	5 FPS	**29.63**	18.43	**37.66**	24.71	26.4
	8 FPS	**29.63**	17.16	36.69	25.88	25.6
	16 FPS	27.41	**19.70**	**37.66**	**29.41**	**27.1**

52.78%, and Qwen3.5-27B achieves the best Animal accuracy of 53.00%. Visually salient motion may contribute to these results. AIGC, Industry, Games, and Human videos show lower performance; synthetic artifacts, subtle movements, fast camera motion, and complex interactions are possible contributors, but their effects are not isolated here. Additional results are provided in Appendix G.3.

4.3 Frame-rate Analysis

To study the effect of temporal sampling density, we conduct a frame-rate ablation on four representative models: Gemini-3.1-Pro, Gemini-3-Flash, Qwen3.5-27B, and Qwen3.5-397B-A17B. For each model, we evaluate four sampling settings: 1 FPS, 5 FPS, 8 FPS, and 16 FPS.

As shown in Table 3, denser sampling generally improves performance, but the gains are model-dependent and not strictly monotonic. Gemini-3.1-Pro improves from 26.9% at 1 FPS to 38.3% at 5 FPS, but does not further improve at 8 or 16 FPS, with 36.3% and 36.6% overall accuracy, respectively. Gemini-3-Flash also benefits substantially, increasing from 26.9% to 37.0% at 8 FPS, but drops to 34.9% at 16 FPS. For open-source models, Qwen3.5-27B shows a gradual gain from 24.1% to 25.7% as the frame rate increases, while Qwen3.5-397B-A17B improves from 22.8% to 26.4% at 5 FPS and further to 27.1% at 16 FPS. For frame-sequence models, the cap rises from 64 to 128 frames above 1 FPS, so rate and budget effects are not isolated. The event-coverage audit in Appendix G.2 confirms that denser sampling covers more annotated event intervals, but errors persist even when all necessary intervals are hit. Thus, interval coverage alone does not guarantee correct answers; the audit does not identify whether remaining errors arise from visual encoding, temporal integration, or answer generation.

4.4 Duration Analysis

We further analyze how accuracy varies with video duration. We select three representative models for this analysis: Seed-2.0-Pro, Gemini-3.1-Pro under 8 FPS sampling, and Qwen3.5-27B under 8 FPS sampling. We use six duration bins: $\leq$5s, 5–10s, 10–20s, 20–60s, 1–3min, and $\geq$3min. As shown in Figure 4, performance generally decreases as videos become longer. Seed-2.0-Pro maintains over 41% accuracy on videos shorter than 20 seconds, but drops to 33.73% on 20–60s videos and 4.76% on videos longer than 3 minutes. Gemini-3.1-Pro shows a similar trend under 8 FPS sampling, decreasing from 40.77% on 5–10s videos to 28.99% on 20–60s videos and 4.76% on the longest videos. Qwen3.5-27B is overall weaker and also drops to 0.00% on the $\geq$3min bucket. This pattern is consistent with greater difficulty locating brief evidence in longer videos, but does not isolate localization from other sources of error. The 5–10s bucket performs comparably to or better than the $\leq$5s bucket; surrounding context is one possible explanation, alongside differences in sample composition. Since the two longest buckets contain relatively few samples, these results should be interpreted as indicative trends. Task and content differences across duration buckets prevent a causal attribution to video length alone. More evaluation results are provided in Appendix G.4.

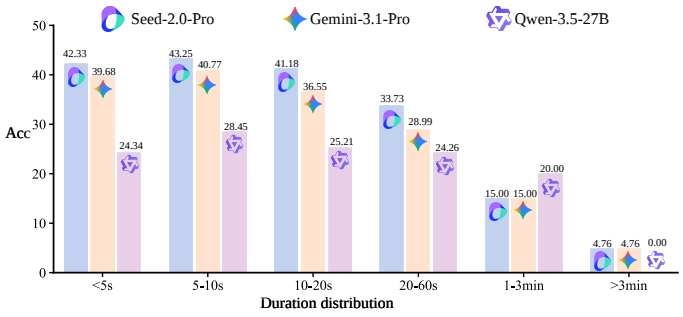


Figure 4: Accuracy of three representative models across video-duration buckets.

5 Conclusion

We introduce Moment-Video, a benchmark of 1,000 human-verified video-QA pairs spanning seven domains and four tasks. Tests on 34 MLLMs reveal difficulty using brief, answer-critical evidence, especially for temporal counting. Denser sampling helps some models, yet errors persist in the audited subset even when frames hit all necessary event intervals. Coverage alone therefore does not ensure correct answers. Lower accuracy on longer videos also suggests localization challenges. These findings motivate video models that better capture, preserve, and use brief visual events.

AI USE STATEMENT

We used generative AI tools to polish the manuscript for grammar, clarity, and readability. All AI-assisted edits were reviewed and verified by the authors for technical accuracy and consistency with the underlying research. As described in Appendix D, we also used Gemini-3-Flash-Preview to screen manually annotated samples for potentially unclear questions, missing referents, ambiguous answers, inconsistent task labels, and possible shortcuts. The model only flagged potential issues; human annotators verified the samples and made the final decisions to accept, revise, or remove them. The screening model did not generate ground-truth answers or determine their final correctness. For evaluation, we used an LLM to assess the semantic equivalence between open-ended model predictions and human-verified reference answers, following the protocol in Section 3.4 and the prompt in Appendix I. The evaluator did not receive the videos and was instructed to base its judgments only on the supplied reference answer and model prediction, without introducing external knowledge. We assessed its agreement with human judgments as reported in Appendix F.

The authors reviewed all AI-assisted contributions for technical accuracy, appropriate attribution, and consistency with the underlying data and experiments. We take responsibility for the final content of this work, including its text, annotations, results, claims, and associated artifacts.

ETHICS STATEMENT

MOMENT-VIDEO is a benchmark for evaluating video understanding and does not involve the development or deployment of systems for high-stakes decision-making. The benchmark is constructed from existing public datasets, publicly accessible videos, and self-collected recordings. We followed the applicable terms of use and licenses when collecting and processing the data. The annotations focus exclusively on visually observable events and do not involve the inference of sensitive personal attributes. We manually reviewed the collected samples and excluded content containing sensitive personal information or other inappropriate material. Any future release of the benchmark will respect the licenses and redistribution conditions of the original video sources.

REPRODUCIBILITY STATEMENT

We provide detailed documentation to support the reproducibility of MOMENT-VIDEO and its evaluation results. Section 3 describes the benchmark construction procedure, event-selection criteria, task taxonomy, and evaluation protocol. The appendices provide the data sources and processing details, dataset statistics, annotation and quality-control procedures, human solvability results, model-specific inference settings, frame-sampling configurations, additional experimental results, and the complete LLM-as-a-Judge prompt. The reported model evaluations use fixed benchmark annotations and explicitly documented input settings.

REFERENCES

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.

Build AI. Egocentric-10K. Hugging Face Datasets, 2025. URL https://huggingface.co/datasets/builddotai/Egocentric-10K.

ByteDance Seed Team. Seed2.0 model card: Towards intelligence frontier for real-world complexity, February 2026. URL https://seed.bytedance.com/zh/seed2. Model Card.

Keshigeyan Chandrasegaran, Agrim Gupta, Lea M. Hadzic, Taran Kota, Jimming He, Cristobal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 53168–53197. Curran Associates, Inc., 2024. doi: 10.52202/079017-1684. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/5f2809607f692d79a01c05c43d702883-Paper-Datasets_and_Benchmarks_Track.pdf.

Dongping Chen, Yue Huang, Siyuan Wu, Jingyu Tang, Liuyi Chen, Yilin Bai, Zhigang He, Chenlong Wang, Huichi Zhou, Yiqiang Li, et al. Gui-world: A dataset for gui-oriented multimodal llm-based agents. *arXiv preprint arXiv:2406.10819*, 2024.

Francesco Di Paco, Luca Burattini, Roberto Gabbrielli, Luca Landi, Francesco Marcelloni, Leonardo Marrazzini, Marco Palumbo, and Marco Pirozzi. Aisafety: An ai-based smart system for enhancing operator safety in production processes. *Safety Science*, 199:107201, 2026.

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. *arXiv preprint arXiv:2503.21776*, 2025.

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 *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 24108–24118, 2025a.

Chaoyou Fu, Haojia Lin, Xiong Wang, Yi-Fan Zhang, Yunhang Shen, Xiaoyu Liu, Yangze Li, Zuwei Long, Heting Gao, Ke Li, et al. Vita-1.5: Towards gpt-4o level real-time vision and speech interaction. *arXiv preprint arXiv:2501.01957*, 2025b.

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

Gemma Team, Google DeepMind. Gemma 4: Our most capable open models to date. `https://blog.google/innovation-and-ai/technology/developers-tools/gemma-4/`, 4 2026.

Google DeepMind. Introducing our most intelligent model yet. with state-of-the-art reasoning to help you learn, build, and plan anything. Google Blog, 2025. URL `https://deepmind.google/models/gemini/`.

Wenyi Hong, Yean Cheng, Zhuoyi Yang, Weihan Wang, Lefan Wang, Xiaotao Gu, Shiyu Huang, Yuxiao Dong, and Jie Tang. Motionbench: Benchmarking and improving fine-grained video motion understanding for vision language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8450–8460, 2025a.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, et al. Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*, 2025b.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos, 2025. URL `https://arxiv.org/abs/2501.13826`.

Hugh GP Hunt. Dataset of photographed lightning events attaching to and around the brixton tower, johannesburg, south africa for the 2015-2016 thunderstorm season. *Data in Brief*, 30:105630, 2020.

Kimi Team, Angang Du, Bohong Yin, Bowei Xing, Bowen Qu, Bowen Wang, Cheng Chen, Chenlin Zhang, Chenzhuang Du, Chu Wei, et al. Kimi-VL technical report. *arXiv preprint arXiv:2504.07491*, 2025.

Jinghang Li, Shichao Li, Qing Lian, Peiliang Li, Xiaozhi Chen, and Yi Zhou. Toward deep representation learning for event-enhanced visual autonomous perception: The eap dataset. *IEEE Transactions on Robotics*, 42:1643–1661, 2026a.

Jingyao Li, Jingyun Wang, Molin Tan, Haochen Wang, Cilin Yan, Likun Shi, Jiayin Cai, Xiaolong Jiang, and Yao Hu. Crossvid: A comprehensive benchmark for evaluating cross-video reasoning in multimodal large language models. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 6244–6252, 2026b.

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 (CVPR)*, pp. 22195–22206, 2024.

Xinhao Li, Ziang Yan, Desen Meng, Lu Dong, Xiangyu Zeng, Yinan He, Yali Wang, Yu Qiao, Yi Wang, and Limin Wang. Videochat-r1: Enhancing spatio-temporal perception via reinforcement fine-tuning. *arXiv preprint arXiv:2504.06958*, 2025.

Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13536–13545, 2021.

Yuanxin Liu, Kun Ouyang, Haoning Wu, Yi Liu, Lin Sui, Xinhao Li, Yan Zhong, Y. Charles, Xinyu Zhou, and Xu Sun. Videoreasonbench: Can mllms perform vision-centric complex video reasoning?, 2026. URL https://arxiv.org/abs/2505.23359.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, pp. 46212–46244. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/90ce332aff156b910b002ce4e6880dec-Paper-Datasets_and_Benchmarks.pdf.

Moonshot AI. Kimi K2.6 tech blog: Advancing open-source coding. Kimi Blog, 2026. URL https://www.kimi.com/en/blog/kimi-k2-6.

Xun Long Ng, Kian Eng Ong, Qichen Zheng, Yun Ni, Si Yong Yeo, and Jun Liu. Animal kingdom: A large and diverse dataset for animal behavior understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 19023–19034, 2022.

OpenAI. GPT-5.4 model. OpenAI API Documentation, 2026. URL https://developers.openai.com/api/docs/models/gpt-5.4. Accessed: 2026-09-25.

Qwen Team. Qwen3.6-27B: Flagship-level coding in a 27B dense model, April 2026a. URL https://qwen.ai/blog?id=qwen3.6-27b.

Qwen Team. Qwen3.6-35B-A3B: Agentic coding power, now open to all, April 2026b. URL https://qwen.ai/blog?id=qwen3.6-35b-a3b.

Qwen Team. Qwen3.5-Omni technical report. *arXiv preprint arXiv:2604.15804*, 2026c. URL https://arxiv.org/abs/2604.15804.

Qwen Team. Qwen3.5: Towards native multimodal agents, February 2026d. URL https://qwen.ai/blog?id=qwen3.5.

Seyed Morteza Safdarnejad, Xiaoming Liu, Lalita Udpa, Brooks Andrus, John Wood, and Dean Craven. Sports videos in the wild (svw): A video dataset for sports analysis. In *2015 11th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, volume 1, pp. 1–7. IEEE, 2015.

Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

Boyuan Sun, Jiaxing Zhao, Xiang Chen, Xihan Wei, and Qibin Hou. Llava-octopus: Unlocking instruction-driven adaptive projector fusion for video understanding. *arXiv preprint arXiv:2501.05067*, 2025.

Nien-En Sun, Yu-Ching Lin, Shao-Ping Chuang, Tzu-Han Hsu, Dung-Ru Yu, Ho-Yi Chung, and Tsì-Uí İk. Tracknetv2: Efficient shuttlecock tracking network. In *2020 International Conference on Pervasive Artificial Intelligence (ICPAI)*, pp. 86–91. IEEE, 2020.

Chongjun Tu, Lin Zhang, Pengtao Chen, Peng Ye, Xianfang Zeng, Wei Cheng, Gang Yu, and Tao Chen. Favor-bench: A comprehensive benchmark for fine-grained video motion understanding. *arXiv preprint arXiv:2503.14935*, 2025.

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

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*, 2025b.

Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems (NeurIPS)*, volume 37, pp. 28828–28857. Curran Associates, Inc., 2024. doi: 10.52202/079017-0907. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/329ad516cf7a6ac306f29882e9c77558-Paper-Datasets_and_Benchmarks_Track.pdf.

XiaomiMiMo. MiMo-V2.5. Hugging Face Model Collection, 2026. URL https://huggingface.co/collections/XiaomiMiMo/mimo-v25.

Anyi Xu, B Li, Bangcai Lin, Bing Xue, BingCheng Xian, Bingzheng Xu, Bochao Wu, Bowei Zhang, Boyi Deng, CC Yu, et al. DeepSeek-V4.1-Flash: Pushing the limits of KV cache compression. *arXiv preprint arXiv:2609.19969*, 2026.

Ziang Yan, Xinhao Li, Yinan He, Zhengrong Yue, Xiangyu Zeng, Yali Wang, Yu Qiao, Limin Wang, and Yi Wang. VideoChat-R1.5: Visual test-time scaling to reinforce multimodal reasoning by iterative perception. *arXiv preprint arXiv:2509.21100*, 2025.

Biao Yang, Bin Wen, Boyang Ding, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, et al. Kwai keye-vl 1.5 technical report. *arXiv preprint arXiv:2509.01563*, 2025.

Chun-Hsiao Yeh, Chenyu Wang, Shengbang Tong, Ta-Ying Cheng, Ruoyu Wang, Tianzhe Chu, Yuexiang Zhai, Yubei Chen, Shenghua Gao, and Yi Ma. Seeing from another perspective: Evaluating multi-view understanding in mllms. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*, volume 40, pp. 12000–12008, 2026.

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*, 2025a.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Llava-video: Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024.

Zheyuan Zhang, Wanying Dou, Linkai Peng, Hongyi Pan, Ulas Bagci, and Boqing Gong. Videoads for fast-paced video understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 21812–21821, October 2025b.

Xiangyu Zhao, Peiyuan Zhang, Kexian Tang, Xiaorong Zhu, Hao Li, Wenhao Chai, Zicheng Zhang, Renqiu Xia, Guangtao Zhai, Junchi Yan, et al. Envisioning beyond the pixels: Benchmarking reasoning-informed visual editing. *Advances in Neural Information Processing Systems (NeurIPS)*, 38, 2026.

Yilun Zhao, Haowei Zhang, Lujing Xie, Tongyan Hu, Guo Gan, Yitao Long, Zhiyuan Hu, Weiyuan Chen, Chuhan Li, Zhijian Xu, et al. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 8475–8489, 2025.

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 13691–13701, June 2025.

A BENCHMARK DATA SOURCES

MOMENT-VIDEO is constructed from diverse video sources to cover momentary visual event across both real-world and virtual scenarios. The data sources include existing open-source video benchmarks, self-collected screen recordings, self-recorded real-world videos, AI-generated videos, and publicly available online videos such as YouTube clips. During data collection, we prioritize videos that contain visually clear momentary actions or short-duration state transitions. We exclude videos whose key events are severely blurred, heavily occluded, or mainly dependent on audio, subtitles, or external context. For videos collected from public online sources, we record the source metadata and access information. Raw videos are released only when redistribution is permitted by the corresponding license or when the videos are self-collected; otherwise, we release annotations, metadata, and source links. Table 4 summarizes the data sources for each fine-grained subcategory in MOMENT-VIDEO. The table is intended to document the provenance of different video types and improve the transparency of benchmark construction.

B EXPERIMENTAL SETUP

We use each model's supported input interface and preprocessing pipeline. Models with native video support are evaluated through their video interfaces; the explicit frame budgets below apply to frame-sequence inputs, not provider-managed video preprocessing. For explicitly sampled inputs, we first extract candidate frames at the specified rate. If their number exceeds the frame budget, we uniformly subsample them across the full video duration rather than truncating the video.

In the main evaluation, frame-sequence models use 1 FPS sampling with at most 64 frames per video, except GPT-5.4, which uses at most 50 frames through the Amazon API. Seed-2.0-Pro, Seed-2.0-Lite, and Seed-2.0-Mini use their native video APIs with default processing. Gemini-3.1-Pro, Gemini-3-Flash, and Gemini-3.1-Flash-Lite use their supported video interfaces with the sampling rate set to 1 FPS. MIMO-v2.5 uses its native video interface with default processing.

Kimi-2.6 and DeepSeek-v4.1-Flash are evaluated through the OpenRouter API. LLaVA-Video-72B, LLaVA-Video-7B, VideoLLaMA3-7B, and VITA-1.5 receive offline-extracted frame sequences. The remaining locally deployed open-source models are served through vLLM, which decodes videos and constructs the sampled frame sequences before forwarding them to the multimodal processor. These explicitly sampled configurations follow the frame budgets specified here.

For the 5, 8, and 16 FPS experiments, frame-sequence models receive at most 128 frames per video; GPT-5.4 retains its 50-frame limit in every evaluated sampling configuration. Native video models continue to use their supported interfaces and provider-specific preprocessing. For short videos, the actual frame count may be below the cap. For longer videos, uniform subsampling can reduce the effective input rate below the nominal candidate sampling rate.

C DATASET STATISTICS

This section provides detailed statistics of **MOMENT-VIDEO**. The benchmark contains 1,000 videos and 1,000 question-answer pairs, with each video corresponding to one question. We report domain and subcategory coverage, task-type composition, answer-format distribution, video duration, source composition, and an event-level temporal audit. These statistics show that MOMENT-VIDEO covers diverse visual scenarios and evaluates multiple levels of momentary visual event understanding.

Domain and Subcategory Distribution. Table 5 reports the number of samples in each fine-grained subcategory. MOMENT-VIDEO covers 7 high-level domains and 25 subcategories, including both real-world scenarios, such as nature, industry, human activities, and animal behaviors, and virtual or synthetic scenarios, such as AIGC videos, GUI recordings, and games.

Task-type Distribution. Table 6 shows the distribution of the four task types in MOMENT-VIDEO. Temporal Occurrence evaluates whether a momentary event occurs, Temporal Counting focuses on counting transient actions or state changes, Action Description requires describing event dynamics, and Temporal Reasoning requires inferring consequences from pre-event, event, and post-event information.

Answer-format Distribution. MOMENT-VIDEO contains both open-ended and multiple-choice questions. We use open-ended questions when the answer is convergent and can be semantically matched with a canonical reference answer. When the answer may be difficult to standardize or could introduce ambiguity in automatic evaluation, we convert the question into a multiple-choice format with manually written distractors. Table 7 reports the distribution of answer formats.

Video Duration and Source Composition. The video durations range from 1.00 to 560.00 seconds, with a mean of 19.25 seconds, a median of 8.02 seconds, and an interquartile range of 5.05–16.34 seconds. In total, 79.0% of videos are shorter than 20 seconds, 41 videos are longer than one minute, and 21 videos are longer than three minutes. The duration buckets contain 190 videos shorter than 5 seconds, 363 videos of 5–10 seconds, 237 videos of 10–20 seconds, 169 videos of 20–60 seconds, 20 videos of 1–3 minutes, and 21 videos longer than 3 minutes. Among the 1,000 videos, 158 (15.8%) are self-recorded or self-collected, including GUI recordings, and 842 (84.2%) come from existing public video sources.

Event Duration and Temporal Position Audit. To quantify the temporal locality of the benchmark events, we conduct a stratified audit of 300 samples covering all domains, subcategories, task types, answer formats, and duration ranges. The audit identifies 609 answer-critical event intervals, including all necessary events for counting and temporal-reasoning questions. The mean and median event durations are 0.703 and 0.540 seconds; 47.9%, 87.7%, and 97.9% of the intervals last no longer than 0.5, 1, and 2 seconds, respectively. A total of 151/300 samples (50.3%) require two or more necessary events. Event centers are distributed across the video timeline rather than concentrated at a fixed location: 14.1%, 20.2%, 22.8%, 23.6%, and 19.2% fall in the 0–20%, 20–40%, 40–60%, 60–80%, and 80–100% relative-position bins, respectively. For an annotated interval $\tau_j = [t_{s,j}, t_{e,j}]$ in a video of duration T, we measure its duration $\Delta_j = t_{e,j} - t_{s,j}$ and relative duration $r_j = \Delta_j / T$. In this audit, 71.4% of necessary event intervals satisfy $r_j \leq 0.10$. These are per-interval proportions, not the union of all necessary intervals in a video. The intervals describe answer-critical visual evidence, which can be shorter than the surrounding action. The duration statistics characterize the audited subset; neither 1 second, 2 seconds, nor 10% is a hard inclusion threshold. Together with the multi-event counts, they show that models must locate and integrate brief evidence at unknown positions.

Task Formulations and Evaluated Capabilities. Given a video V and question q, a model predicts an answer $\hat{y} = f(V, q)$. Let $\mathcal{E}_q = \{\tau_j\}_{j=1}^{m}$ denote the necessary evidence intervals identified for an audited question. These intervals serve the coverage analysis and are not supplied to the model; the full benchmark uses weak temporal grounding during construction rather than dense interval annotations for every sample. The question determines which event, entities, and temporal scope are relevant.

For **TO**, the answer identifies the presence, absence, or immediate change of a queried event or state. For a yes/no occurrence question, this reduces to $y = \mathbf{1}[e \text{ occurs in } V]$; other TO questions name the observed change. The capability is noticing and identifying brief evidence, rather than counting repetitions. For **TC**, $y = |\mathcal{O}_q|$, where $\mathcal{O}_q$ is the set of distinct occurrences, entities performing the specified action, or state changes requested by q. The capability is temporal tracking and quantification without omissions or double counting. For **AD**, $y = d_q(\mathcal{E}_q)$ describes the requested visible event attributes, such as actor, direction, trajectory, target, or interaction. The capability is retaining event-specific dynamic details. For **TR**, $y = g_q(s_{\text{pre}}, \mathcal{E}_q, s_{\text{post}})$ is a consequence inferred by relating the initial state, necessary events, and resulting state. The capability is integrating evidence across time: counting observed departures is TC, whereas inferring how many objects remain from the initial count and departures is TR. These are semantic task targets; each can be expressed as either a multiple-choice or an open-ended question under the scoring protocol in Section 3.4.

D DATA QUALITY CONTROL

We apply a multi-stage quality control process to ensure that each sample is visually grounded, event-centric, and evaluable. All samples are manually annotated and verified by two annotators. After the initial annotation, we use Gemini-3-Flash-Preview only as a screening tool to flag potentially unclear

Table 4: Data sources of the fine-grained subcategories in **MOMENT-VIDEO**.

Domain	Subcategory	Source
AIGC	Artifacts	Publicly available subset from RISEBench (Zhao et al., 2026).
GUI	XR	Publicly available subset from GUI-World (Chen et al., 2024).
GUI	Website, Application	Publicly available subset from GUI-World (Chen et al., 2024) and self-collected screen recordings.
GUI	Operating System	Self-collected screen recordings.
Nature	Lightning	Publicly available video dataset subset (Hunt, 2020) and YouTube clips.
Nature	Other Nature Phenomena	YouTube clips.
Industry	Manual Assembly	Publicly available subset from Egocentric-10K (Build AI, 2025).
Industry	Industrial Machinery	YouTube clips.
Games	FPS	YouTube gameplay clips from CS2 and Valorant.
Games	MOBA	YouTube gameplay clips from League of Legends.
Games	Combat	YouTube gameplay clips from Street Fighter 6, Tekken 8, and Guilty Gear.
Games	Music	YouTube gameplay clips from osu!, Phigros, and Muse Dash.
Human	Basketball, Soccer, Volleyball	Publicly available subsets from MultiSports (Li et al., 2021) and SVW (Safdarnejad et al., 2015).
Human	Other Ball Sports, Non-ball Sports, Daily Life	Publicly available subsets from MultiSports (Li et al., 2021), SVW (Safdarnejad et al., 2015), the shuttlecock trajectory dataset(Sun et al., 2020) and UCF101 (Soomro et al., 2012).
Animal	Birds, Mammals, Reptiles, Amphibians, Fish, Invertebrates	Publicly available subset from Animal Kingdom (Ng et al., 2022).

questions, missing referents, ambiguous answers, inconsistent task labels, or shortcut-prone samples. The screening model is not used to generate ground-truth answers or determine final correctness.

The annotators then conduct human cross-checking over the target event, question wording, reference answer, task label, domain label, subcategory label, and answer format. They verify that the question is grounded in the selected momentary event, that the answer is uniquely determined by visual evidence, and that the sample does not primarily rely on audio, subtitles, common sense, external knowledge, or persistent textual shortcuts. For multiple-choice questions, annotators additionally check that exactly one option is correct and that distractors are visually plausible.

Samples are revised or removed when the target event is not clearly visible, the question can be answered without the critical moment, the answer is ambiguous, or the distractors introduce unintended shortcuts.

E HUMAN SOLVABILITY STUDY

To further validate the quality of MOMENT-VIDEO, we conduct a human solvability study on examples sampled from the benchmark. We initially recruit 20 participants, and each participant answers 20 questions after watching the corresponding videos. Participants do not have access to the reference answers or model predictions. After quality filtering, we retain 15 valid questionnaires, resulting in 300 valid human responses. We compare each human response with the reference answer to evaluate whether the questions are answerable by human viewers.

As shown in Table 8, human participants achieve an overall accuracy of 84.33% on the valid responses. This result indicates that most sampled questions in MOMENT-VIDEO are answerable by human viewers after watching the corresponding videos. Meanwhile, the accuracy is still far from saturated, suggesting that momentary visual event understanding remains non-trivial even for humans and requires careful observation of brief visual evidence.

Table 5: **MOMENT-VIDEO** domain and subcategory stats (one QA per video).

Domain (#)	Subcategory	#Videos / #QA
AIGC (79)	Artifacts	79
GUI (218)	XR	8
	Website	97
	Application	91
	Operating System	22
Nature (72)	Lightning	49
	Other Natural Phenomena	23
Industry (121)	Manual Assembly	96
	Industrial Machinery	25
Games (160)	FPS	60
	MOBA	50
	Combat	30
	Music	20
Human (250)	Basketball	57
	Soccer	38
	Volleyball	40
	Other Ball Sports	37
	Non-ball Sports	40
	Daily Life	38
Animal (100)	Birds	42
	Mammals	13
	Reptiles	29
	Amphibians	6
	Fish	2
	Invertebrates	8
Total	–	**1,000**

Table 6: Task-type statistics of **MOMENT-VIDEO**.

Task Type	Abbr.	#Questions
Temporal Occurrence	TO	135
Temporal Counting	TC	472
Action Description	AD	308
Temporal Reasoning	TR	85
Total	–	**1,000**

Overall, the human solvability study provides additional evidence that MOMENT-VIDEO is generally human-solvable and reasonably well specified, while still retaining meaningful difficulty due to the brief and localized nature of momentary visual events. This study validates question solvability and answer clarity on a sampled subset; it is not intended as a population-level human performance ceiling for the full benchmark.

F LLM-AS-JUDGE VALIDATION

To validate the reliability of our LLM-as-Judge protocol for open-ended questions, we conduct a human validation study on the same set of 764 open-ended examples in MOMENT-VIDEO. Specifically, we use the predictions from Gemini-3-Flash as the validation set and ask three human annotators to independently judge whether each model prediction is semantically consistent with the reference answer. During this process, annotators only see the model answer and the reference answer; they do not have access to the video, model identity, or the LLM-as-Judge decision. We treat each human annotator's judgment as the reference label and evaluate the LLM-as-Judge decision as a

Table 7: Answer-format statistics of **MOMENT-VIDEO**.

Answer Format	#Questions
Open-ended	764
Multiple-choice	236
Total	**1,000**

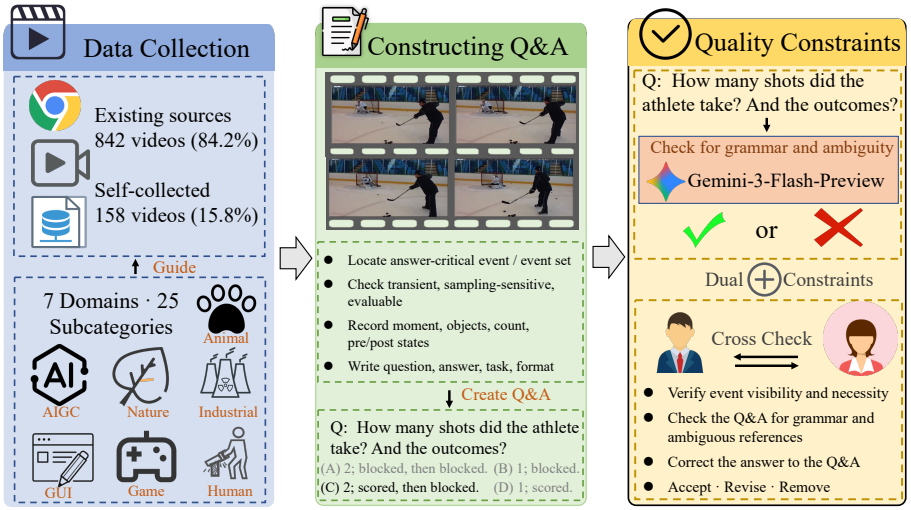


Figure 5: Dataset construction pipeline. Videos are collected under the 7-domain, 25-subcategory taxonomy; annotators then localize answer-critical events before constructing questions and answers. Gemini-3-Flash-Preview is used only to flag potential issues, while two annotators make the final accept, revise, or remove decision.

binary classification result, where positive indicates that the model answer is judged as correct or semantically consistent.

Since the validation labels are imbalanced, with incorrect answers forming the majority of samples, raw accuracy alone may not fully reflect judge reliability. We therefore report multiple metrics, including accuracy, F1 score, false positive rate (FPR), and Cohen's kappa. Accuracy measures overall agreement between each human annotator and the LLM-as-Judge. F1 score evaluates consistency on the positive class, i.e., answers judged as semantically correct. FPR is particularly important in our setting because false positives correspond to incorrect model answers being accepted as correct, which may overestimate model performance. Cohen's kappa measures agreement beyond chance and provides a more conservative estimate under imbalanced label distributions.

As shown in Table 9, the LLM-as-Judge shows strong agreement with human annotations on the same 764 open-ended examples, achieving 96.73%–97.51% accuracy across the three annotators. The F1 score ranges from 88.89% to 91.24%, indicating strong consistency on answers judged as semantically correct. Cohen's kappa ranges from 0.870 to 0.898, showing strong agreement beyond chance even under the imbalanced label distribution.

Importantly, the false positive rate remains very low, ranging from 1.07% to 2.14%. This suggests that the judge rarely accepts incorrect model responses as correct, reducing the risk of overestimating model performance. Overall, these results indicate that our LLM-as-Judge protocol is well aligned with human judgments and provides a reliable automatic evaluation procedure for open-ended answers in MOMENT-VIDEO.

Table 8: Human solvability study on sampled examples from MOMENT-VIDEO.

Metric	Value
Initial participants	20
Valid participants	15
Questions per valid participant	20
Valid human responses	300
Overall human accuracy	84.33%
Participant accuracy range	70.00%–95.00%

Table 9: Validation of LLM-as-Judge against human annotations on the same 764 open-ended Gemini-3-Flash predictions. Cohen's kappa measures agreement beyond chance.

Annotator	#Samples	Acc.(%)	F1(%)	FPR(%)	Kappa
Annotator 1	764	96.73	88.89	2.14	0.870
Annotator 2	764	97.12	90.00	1.53	0.883
Annotator 3	764	97.51	91.24	1.07	0.898
Average	764	97.12	90.04	1.58	0.884

G ADDITIONAL EXPERIMENTAL RESULTS

G.1 RESULTS OF MOMENT-VIDEO UNDER AN 8 FPS INPUT RATE

Table 10 reports additional evaluations at 8 FPS. Frame-sequence inputs are capped at 128 frames per video, except GPT-5.4, which remains capped at 50; native video models use their supported interfaces. Candidate frames exceeding the applicable cap are uniformly subsampled across the full video duration. Thus, 8 FPS is the candidate sampling rate, and the effective input rate can be lower for longer videos. Gray rows show these additional results; non-gray rows retain the main evaluation settings. Bold and underlined values denote the best and second-best results within each model group, respectively.

The updated 8 FPS results show model-dependent gains. Gemini-3-Flash and Gemini-3.1-Pro reach 37.0% and 36.3% overall accuracy, respectively. Among open-source models, DeepSeek-v4.1-Flash achieves 29.0%, followed by Qwen3.5-122B-A10B (27.4%), Qwen3.6-27B (27.3%), Qwen3.5-397B-A17B (25.6%), and Qwen3-VL-235B-A22B (25.5%). Improvements are not universal: Keye-VL-1.5-8B decreases from 15.6% to 13.8%, and VideoLLaMA3-7B from 11.8% to 9.6%. For frame-sequence models other than GPT-5.4, the frame budget also increases from 64 to 128, so their comparisons with 1 FPS do not isolate sampling rate alone.

However, 8 FPS sampling does not fully solve the challenge of momentary visual event understanding. Even with denser temporal inputs, accuracy remains low on tasks involving event detection, counting, description, and reasoning. These results show that increasing the sampling rate alone is insufficient for reliable performance; they do not identify a specific internal cause of the remaining errors.

G.2 EVENT-LEVEL INPUT COVERAGE AUDIT

We audit 300 stratified samples containing 609 necessary event intervals, including every necessary event for counting and temporal-reasoning questions. Sampling is independent of model correctness. We reconstruct the final input-frame timestamps under each model's evaluation sampling rules. An event is hit only if an input timestamp lies within its annotated interval, with no temporal tolerance. A sample is *Full* if all necessary events are hit, *Partial* if only some are hit, and *Miss* if none is hit. Event hit rate uses 609 events as its denominator; Full, Partial, and Miss counts partition the 300 samples.

As shown in Table 11, event hit rates increase from 54.8%–55.3% at 1 FPS to 93.3%–99.2% at 8 FPS. At 1 FPS, Full-group accuracy exceeds the combined Partial+Miss accuracy for all five models. However, Partial and Miss accuracies need not be ordered: the groups differ in task composition, and context outside strictly annotated intervals can provide indirect clues. At 8 FPS, 289–298 of

300 samples achieve Full coverage for the four open-weight models, yet their Full-group accuracy remains 21.5%–30.9%. This distinguishes missing interval coverage from errors that persist when every annotated interval contains a sampled timestamp. A timestamp hit does not guarantee that all relevant event dynamics are resolved. The audit does not attribute remaining errors to a specific internal mechanism.

G.3 Subcategory-level Results

The main paper reports category-level accuracy over seven high-level domains. To provide a more fine-grained view, we further report subcategory-level results over the 25 subcategories in **MOMENT-VIDEO**. Table 12 summarizes the average accuracy of all evaluated models on each subcategory, and Figure 6 visualizes the full model-by-subcategory accuracy matrix.

The subcategory-level results show that model performance varies substantially across fine-grained scenarios, even within the same high-level domain. In general, models achieve relatively higher average accuracy on several animal and nature-related subcategories. For example, the average accuracy reaches 45.52% on *Lightning*, 35.08% on *Birds*, 34.69% on *Reptiles*, and 32.58% on *Mammals*. These subcategories often contain visually salient motion or state changes, such as flying, jumping, lighting variation, or clear object movement, which may be easier to capture when the decisive frames are sampled.

In contrast, several subcategories remain particularly challenging. The average accuracy is low on *Daily Life* (8.59%), *Music* (10.00%), *Artifacts* (11.91%), *Volleyball* (12.57%), and *Soccer* (13.51%). These scenarios often involve subtle state changes, small objects, fast camera motion, dense visual layouts, or synthetic visual artifacts, making the momentary event harder to distinguish from background dynamics. Industrial subcategories are also difficult, with 16.33% on *Manual Assembly* and 13.53% on *Industrial Machinery*, suggesting that current models struggle with small-region safety-related actions and brief mechanical movements.

The results also reveal large variation within the same domain. For example, within GUI videos, *Operating System* achieves 31.29% average accuracy, while *Website* and *Application* remain much lower at 17.32% and 15.11%, respectively. Similarly, within human activities, *Non-ball Sports* reaches 24.78%, whereas *Daily Life* is only 8.59%. This indicates that coarse category-level accuracy can hide important differences among fine-grained scenarios. We note that several subcategories contain relatively few samples, such as *Fish*, *Amphibians*, *XR*, and *Invertebrates*; their average accuracies should therefore be interpreted as indicative rather than conclusive.

Overall, the subcategory-level analysis complements the main results by showing that momentary visual event understanding is highly scenario-dependent and that fine-grained evaluation is necessary for revealing model weaknesses hidden by broad domain-level metrics.

G.4 Duration-wise Results for All Models

In addition to the duration analysis of three representative models in the main paper, we provide model-level duration-wise results for all evaluated models. Figure 7 shows the accuracy of each model across six video-duration buckets: $\leq$5s, 5–10s, 10–20s, 20–60s, 1–3min, and $\geq$3min. Each subplot corresponds to one model, and each bar represents the accuracy within a specific duration bucket.

Overall, the model-level results are consistent with the general trend observed in the main paper. Most models achieve relatively better performance on short-to-medium videos, especially in the 5–10s and 10–20s buckets, while performance generally drops on longer videos. The reported average accuracy reaches 21.51% on 5–10s videos and 21.99% on 10–20s videos, compared with 18.47% on videos shorter than 5 seconds. Surrounding context may contribute to this pattern, but the duration groups also differ in task and content composition. Performance decreases as videos become longer. The average accuracy drops to 16.95% on 20–60s videos, 11.97% on 1–3min videos, and 4.18% on videos longer than 3 minutes. This pattern is consistent with greater difficulty locating relevant evidence in longer videos. However, these comparisons do not directly measure temporal localization or internal aggregation, and do not isolate duration from differences in sample composition. We also observe that the degradation pattern varies across models. Some models remain relatively stable across short- and medium-duration buckets, while others are more sensitive to duration changes. However, the

longest-duration bucket remains consistently difficult for most models. Since the two longest buckets contain relatively few samples, their results should be interpreted as indicative trends rather than definitive conclusions. Overall, these findings establish an association between video duration and answer accuracy on this benchmark, without identifying the internal mechanism responsible for the performance differences.

H PARTIAL MODEL TEST RESULTS

We provide partial model test results in Figures 8, 9, 10, and 11. These examples show representative outputs of different models on MOMENT-VIDEO.

I LLM-AS-JUDGE PROMPT

MOMENT-VIDEO contains both multiple-choice and open-ended questions, and we adopt different evaluation strategies for these two formats. All multiple-choice questions have exactly one correct option. A question with K options is evaluated under K shuffled option orders (e.g., four evaluations for four options). The model response is normalized and matched to the ground-truth option under each order. A sample is counted as correct only if the model selects the semantically correct option in every evaluation.

For open-ended questions, exact string matching is often too strict, since semantically equivalent answers may appear in different surface forms. Therefore, we use an LLM-as-Judge protocol only for open-ended questions. The judge compares the model prediction with the canonical reference answer and determines whether they are semantically consistent. The judge does not see the video, does not re-answer the question, and is not allowed to introduce external knowledge. It is used solely for answer-equivalence checking. All judge calls are performed with temperature 0, and the judge is required to output a structured JSON object. As shown in Figure 12, the judge only evaluates semantic consistency between the model prediction and the reference answer.

Table 10: Main results on **MOMENT-VIDEO**. The **bold** and underlined numbers indicate the best and second-best performance within each group. Gray rows indicate additional 8 FPS results. Frame-sequence inputs use at most 128 frames (GPT-5.4: 50); native video models use their supported interfaces.

Model	Input Settings	Task Type(%)				Video Category(%)							Overall(%)
		TO	TC	AD	TR	AIGC	GUI	Nature	Industry	Games	Human	Animal	
Proprietary Models													
Seed-2.0-Pro	default	**50.37**	**31.14**	47.08	42.35	**45.57**	33.94	**69.44**	**41.32**	**39.37**	27.20	55.00	**39.6**
Gemini-3-Flash	8fps	40.00	27.97	**49.03**	38.82	37.97	**37.61**	56.94	26.45	35.63	**28.00**	58.00	37.0
Gemini-3.1-Pro	8fps	40.00	27.75	45.45	**44.71**	35.44	34.86	58.33	38.02	35.00	23.60	56.00	36.3
Seed-2.0-Lite	default	34.81	25.64	36.04	40.00	22.78	27.98	51.39	30.58	25.62	26.80	52.00	31.3
Gemini-3.1-Flash-Lite	8fps	34.81	24.79	37.34	38.82	27.85	27.98	55.56	29.75	25.00	21.20	**60.00**	31.2
GPT-5.4	8fps	32.59	21.19	40.26	35.29	22.78	34.40	41.67	19.83	29.38	20.00	54.00	29.8
Seed-2.0-Mini	default	32.59	22.88	33.77	25.88	24.05	22.94	58.33	29.75	22.50	20.80	43.00	27.8
Gemini-3.1-Pro	1fps	32.59	19.70	33.44	34.12	27.85	27.52	44.44	18.18	24.37	19.20	46.00	26.9
Gemini-3-Flash	1fps	41.48	18.43	31.82	32.94	**45.57**	20.64	43.06	15.70	20.62	23.20	47.00	26.9
Gemini-3.1-Flash-Lite	1fps	34.07	18.64	30.52	27.06	27.85	22.02	41.67	22.31	20.00	14.00	57.00	25.1
GPT-5.4	1fps	21.48	12.08	30.19	29.41	6.33	23.85	40.28	4.96	18.75	18.00	37.00	20.4
Open-source Models													
DeepSeek-v4.1-Flash	8fps	36.30	19.49	**38.96**	**34.12**	26.58	**40.83**	51.39	17.36	19.38	**20.80**	39.00	**29.0**
Qwen3.5-122B-A10B	8fps	**39.26**	19.49	34.42	27.06	**39.24**	23.39	38.89	**23.14**	21.25	20.40	51.00	27.4
Qwen3.6-27B	8fps	28.89	21.40	34.74	30.59	24.05	26.15	45.83	**23.14**	**23.75**	19.20	50.00	27.3
Qwen3.5-397B-A17B	8fps	29.63	17.16	36.69	25.88	18.99	24.31	50.00	13.22	20.00	19.60	**55.00**	25.6
Qwen3-VL-235B-A22B	8fps	16.30	**22.67**	33.12	28.24	5.06	26.61	50.00	21.49	22.50	19.60	46.00	25.5
Qwen3.5-27B	8fps	22.22	19.92	36.36	20.00	12.66	24.31	52.78	20.66	22.50	15.20	53.00	25.3
Kimi-2.6	1fps	25.93	15.89	36.69	30.59	15.19	37.16	41.67	12.40	16.25	20.40	34.00	24.9
Qwen3.6-35B-A3B	8fps	28.15	19.49	32.47	21.18	21.52	22.48	40.28	21.49	20.62	18.40	48.00	24.8
Gemma-4-31B	8fps	32.59	17.16	31.17	27.06	26.58	16.97	47.22	17.36	**23.75**	15.60	54.00	24.4
Qwen3.5-27B	1fps	20.74	19.07	32.47	27.06	11.39	21.56	41.67	18.18	18.75	20.00	53.00	24.1
Qwen3.5-35B-A3B	8fps	19.26	17.58	36.04	21.18	8.86	23.85	**55.56**	17.36	16.25	18.00	47.00	23.8
Qwen3.5-9B	8fps	34.81	15.89	31.17	20.00	35.44	22.02	50.00	12.40	13.75	15.20	48.00	23.5
Qwen3.5-397B-A17B	1fps	27.41	14.83	32.79	23.53	16.46	23.39	47.22	18.18	17.50	14.80	43.00	22.8
DeepSeek-v4.1-Flash	1fps	30.37	14.62	28.90	31.76	27.85	32.57	41.67	9.92	12.50	17.20	28.00	22.6
Qwen3.6-27B	1fps	19.26	17.16	31.82	23.53	10.13	26.61	41.67	19.01	16.25	14.40	44.00	22.5
MIMO-v2.5	default	18.52	17.16	30.84	23.53	2.53	24.31	52.78	19.01	19.38	14.80	37.00	22.1
Qwen3.5-122B-A10B	1fps	25.19	14.19	27.60	23.53	18.99	22.02	40.28	17.36	13.75	14.40	35.00	20.6
Qwen3.6-35B-A3B	1fps	17.04	14.41	29.22	24.71	3.80	18.35	41.67	22.31	15.62	15.20	39.00	20.2
Qwen3.5-35B-A3B	1fps	20.00	16.10	26.30	18.82	8.86	19.27	45.83	19.83	13.75	12.80	40.00	20.0
Gemma-4-31B	1fps	30.37	13.56	24.35	22.35	27.85	11.93	41.67	17.36	15.62	15.60	36.00	19.9
Qwen3-VL-30B-A3B	8fps	15.56	15.47	26.30	22.35	7.59	16.51	48.61	15.70	18.12	13.20	36.00	19.4
InternVL3.5-241B-A28B	8fps	18.52	14.41	25.65	20.00	6.33	12.84	45.83	7.44	20.60	15.20	43.00	18.9
Qwen3.5-4B	8fps	17.04	17.58	23.05	11.76	7.59	15.60	44.44	14.05	13.13	14.80	40.00	18.7
Qwen3.5-9B	1fps	18.52	13.35	26.62	18.82	10.13	17.43	40.28	12.40	11.87	14.80	40.00	18.6
Qwen3-VL-8B	8fps	20.00	15.04	22.73	18.82	16.46	16.06	41.67	11.57	11.25	13.20	41.00	18.4
InternVL3.5-30B-A3B	8fps	17.78	14.19	23.70	16.47	6.33	12.84	44.44	19.10	16.88	12.80	31.00	17.8
InternVL3.5-8B	8fps	14.81	16.10	20.78	16.47	2.53	10.55	41.67	16.53	16.88	13.60	38.00	17.4
Qwen3.5-4B	1fps	19.26	13.77	21.43	17.65	8.86	16.97	38.89	13.22	15.00	11.20	32.00	17.2
InternVL3.5-4B	8fps	15.56	15.89	21.43	10.59	5.06	11.47	31.94	14.88	15.62	13.60	42.00	17.1
Qwen3-VL-235B-A22B	1fps	15.56	12.08	23.38	23.53	5.06	16.97	38.89	10.74	10.62	14.00	36.00	17.0
InternVL3.5-30B-A3B	1fps	12.59	16.10	21.75	11.76	3.80	10.55	38.89	15.70	15.00	18.40	27.00	17.0
Qwen3-VL-30B-A3B	1fps	18.52	13.14	20.45	20.00	8.86	14.22	43.06	14.05	15.62	9.60	32.00	16.7
InternVL3.5-8B	1fps	15.56	15.04	20.13	15.29	3.80	11.93	38.89	13.22	13.75	14.00	37.00	16.7
LLaVA-Video-72B	1fps	13.33	15.04	18.83	17.65	0.00	8.26	34.72	19.83	16.25	15.20	31.00	16.2
InternVL3.5-241B-A28B	1fps	19.26	11.02	20.45	18.82	1.27	12.39	40.28	9.09	11.87	13.60	36.00	15.7
Keye-VL-1.5-8B	1fps	14.81	12.92	20.45	14.12	0.00	10.09	29.17	17.36	14.37	14.40	33.00	15.6
GLM-4.6V	8fps	14.81	9.32	21.75	20.00	2.53	11.47	36.11	9.92	10.62	13.60	34.00	14.80
InternVL3.5-4B	1fps	14.81	13.14	17.86	7.06	3.80	10.55	29.17	14.05	13.75	10.80	30.00	14.3
Qwen3-VL-4B	8fps	15.56	10.38	19.48	14.12	3.80	11.47	44.44	9.92	9.38	8.80	33.00	14.2
GLM-4.6V	1fps	15.56	8.69	19.81	21.18	5.06	10.55	31.94	7.44	11.25	12.40	33.00	14.1
Qwen3-VL-8B	1fps	15.56	10.59	17.86	16.47	8.86	13.30	41.67	9.92	6.88	8.80	29.00	14.0
Qwen3-VL-4B	1fps	14.81	8.69	19.48	21.18	1.27	14.22	41.67	7.44	9.38	9.20	30.00	13.9
Keye-VL-1.5-8B	8fps	15.56	11.44	17.86	9.41	2.53	7.80	31.94	12.40	10.62	14.80	27.00	13.8
LLaVA-Video-7B	8fps	11.11	12.92	15.26	10.59	1.27	8.26	31.94	12.40	16.88	10.80	21.00	13.2
LLaVA-Video-7B	1fps	11.11	13.14	13.96	7.06	0.00	8.72	23.61	8.26	15.62	12.80	23.00	12.6
VideoLLaMA3-7B	1fps	7.52	13.69	11.72	8.33	0.00	10.44	25.35	4.96	10.14	15.64	15.00	11.8
GLM-4.6V-Flash	8fps	12.59	6.99	15.26	18.82	0.00	8.62	45.83	14.05	10.00	6.00	14.00	11.3
GLM-4.6V-Flash	1fps	11.11	7.20	14.29	20.00	1.27	8.72	43.06	13.22	10.00	6.00	12.00	11.0
VITA-1.5	1fps	19.26	8.49	13.42	10.59	0.00	5.31	25.00	8.26	11.25	8.80	26.00	10.6
VideoLLaMA3-7B	8fps	8.89	11.44	8.12	5.88	0.00	4.59	22.22	7.44	11.87	12.40	11.00	9.6

Table 11: Input coverage audit on 300 samples. Group cells report sample count / accuracy (%). Hit rate is measured over 609 events; overall accuracy is measured over the audited subset. A dash indicates an empty group.

Model	FPS	Event hit (%)	Full: n / Acc.	Partial: n / Acc.	Miss: n / Acc.	Overall (%)
GPT-5.4	1	54.8	118 / 32.2	94 / 7.4	88 / 25.0	22.3
GPT-5.4	8	93.3	271 / 33.9	21 / 14.3	8 / 37.5	32.7
Qwen3.5-27B	1	55.3	119 / 30.3	93 / 21.5	88 / 25.0	26.0
Qwen3.5-27B	8	99.2	298 / 27.5	2 / 0.0	0 / –	27.3
Qwen3.6-27B	1	55.3	119 / 28.6	93 / 17.2	88 / 27.3	24.7
Qwen3.6-27B	8	99.2	298 / 30.9	2 / 0.0	0 / –	30.7
Gemma-4-31B	1	55.3	119 / 27.7	93 / 10.8	88 / 20.5	20.3
Gemma-4-31B	8	96.6	289 / 26.0	8 / 0.0	3 / 33.3	25.3
InternVL3.5-30B-A3B	1	55.3	119 / 29.4	93 / 14.0	88 / 20.5	22.0
InternVL3.5-30B-A3B	8	96.6	289 / 21.5	8 / 0.0	3 / 0.0	20.7

Table 12: Average subcategory-level accuracy of all models on MOMENT-VIDEO.

Domain	**Subclass**	**#Samples**	**Avg. Acc.(%)**	**Domain**	**Subclass**	**#Samples**	**Avg. Acc.(%)**
AIGC	Artifacts	79	11.91	Games	Music	20	10.00
GUI	XR	8	26.52	Human	Basketball	57	16.84
GUI	Website	97	17.32	Human	Soccer	38	13.51
GUI	Application	91	15.11	Human	Volleyball	40	12.57
GUI	Operating System	22	31.29	Human	Other Ball Sports	37	14.00
Nature	Lightning	49	45.52	Human	Non-ball Sports	40	24.78
Nature	Other Natural Phenomena	23	30.95	Human	Daily Life	38	8.59
Industry	Manual Assembly	96	16.33	Animal	Birds	42	35.08
Industry	Industrial Machinery	25	13.53	Animal	Mammals	13	32.58
Games	FPS	60	13.70	Animal	Reptiles	29	34.69
Games	MOBA	50	17.06	Animal	Amphibians	6	24.51
Games	Combat	30	22.84	Animal	Fish	2	73.53
Animal	Invertebrates	8	51.84				

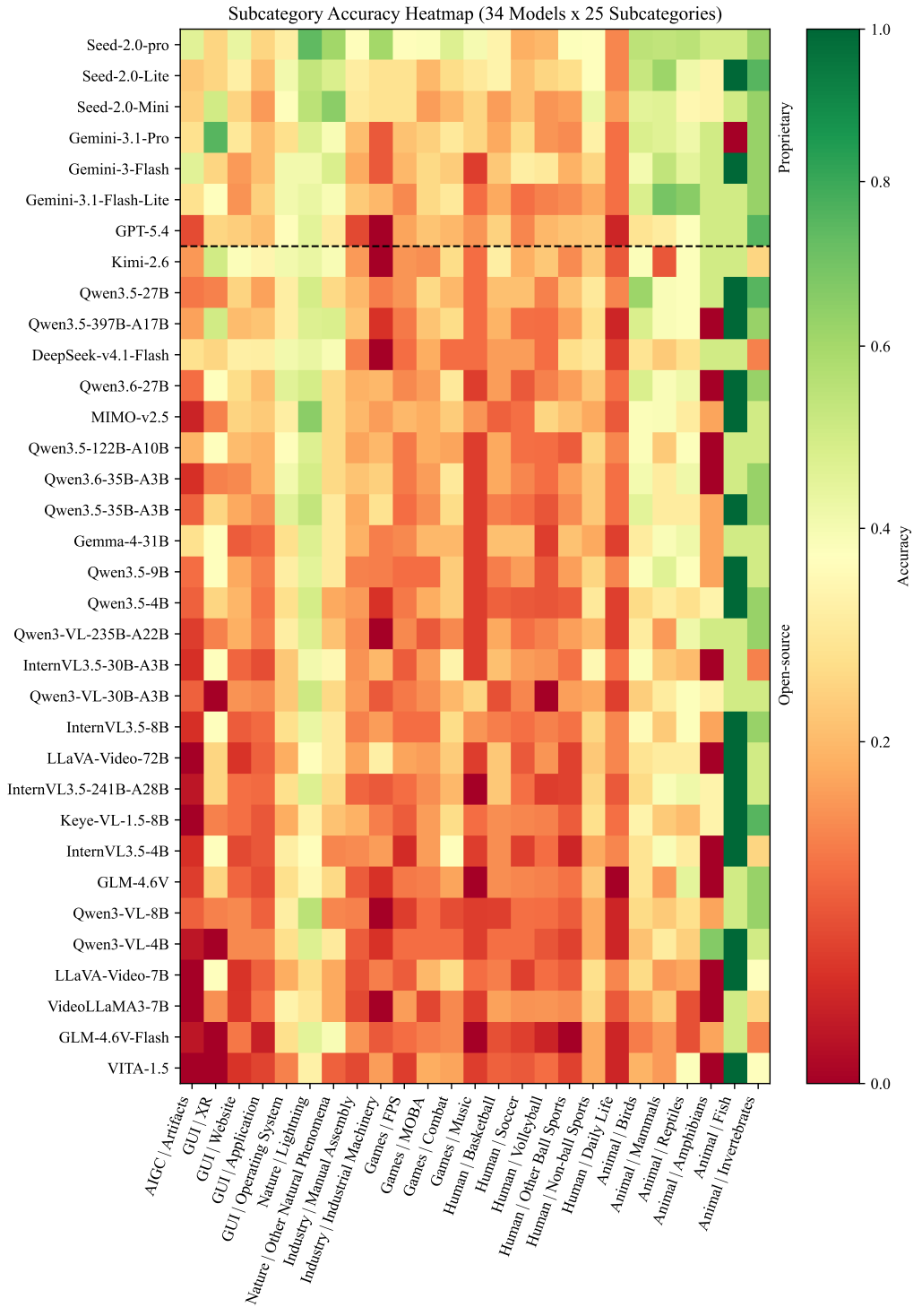
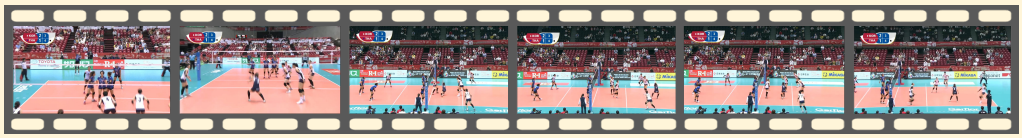


Figure 6: Subcategory-level accuracy heatmap on **MOMENT-VIDEO**. Rows correspond to 34 evaluated models and columns correspond to 25 fine-grained subcategories. The color indicates accuracy, and the dashed horizontal line separates proprietary and open-source models.



Figure 7: Duration-wise accuracy of all evaluated models on **MOMENT-VIDEO**. Each subplot corresponds to one model, and bars represent accuracy across six video-duration buckets. The results show that most models perform better on short-to-medium videos and degrade when rapid events are embedded in longer temporal contexts.

Human



- Subcategory: Volleyball
- Task Type: Temporal Counting
- Q: How many times did right-side team No. 8 hit the ball?
- GT: 4 times.

Model Answer:

- Seed-2.0-pro: The right-side (Korea) team's No. 8 hit the ball 3 times.
- Gemini-3.1-Pro: The action appears a total of 4 times.
- GPT-5.4: 2.
- Kimi-2.6: 2.
- Qwen3.5-27B: The right-side team No. 8 hit the ball 2 times.
- MIMO-v2.5: Player No. 8 does not make contact with the ball at any point.

Animal



- Subcategory: Mammals
- Task Type: Temporal Counting
- Q: What is the maximum number of deer visible in the video at the same time?
- GT: 4 deer.

Model Answer:

- Seed-2.0-pro: The maximum number of deer visible at the same time in the video is 3: 1 adult deer, plus 2 fawns.
- Gemini-3.1-Pro: The video shows a maximum of 3 deer at the same time.
- GPT-5.4: 3 deer.
- Kimi-2.6: 3.
- Qwen3.5-27B: The maximum number of deer visible in the video at the same time is three.
- MIMO-v2.5: Based on the video, the maximum number of deer visible at the same time is 3.

Figure 8: Partial model test results on **MOMENT-VIDEO** (Part I).

GUI

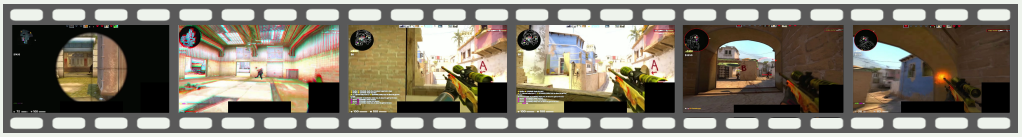


- Subcategory: Website
- Task Type: Action Description
- Q: Which button was clicked to navigate to the next page? Give the button name.
- GT: NeurIPS.

Model Answer:

- Seed-2.0-pro: The clicked link (navigation target) is NeurIPS (the NeurIPS venue entry in the All Venues list, clicking it triggers navigation to the new page).
- Gemini-3.1-Pro: The user clicked on the Leibniz-Centre General Linguistics button.
- GPT-5.4: L4DC.
- Kimi-2.6: refresh.
- Qwen3.5-27B: The button clicked to navigate to the next page was "Show all 758 venues".
- MIMO-v2.5: Based on the video, the user clicks on the button labeled 2026.

Games



- Subcategory: FPS
- Task Type: Temporal Counting
- Q: How many shots were fired using the sniper rifle in the video?
- GT: A total of 6 sniper shots were fired in the video.

Model Answer:

- Seed-2.0-pro: A total of 6 sniper rifle (AWP) shots are fired in the video.
- Gemini-3.1-Pro: The action appears a total of 5 times.
- GPT-5.4: 6 shots.
- Kimi-2.6: Total: 4 shots.
- Qwen3.5-27B: In the video, the player fires the sniper rifle twice.
- MIMO-v2.5: The sniper rifle is fired 4 times in the video.

Figure 9: Partial model test results on **MOMENT-VIDEO** (Part II).

Industrial



- Subcategory: Manual Assembly
- Task Type: Temporal Counting
- Q: Analyze this video. How many plastic caps were removed in the video?
- GT: In the video, two plastic caps were removed.

Model Answer:

- Seed-2.0-pro: 3 plastic caps were removed in the video.
- Gemini-3.1-Pro: The video shows a user removing 1 plastic caps.
- GPT-5.4: 1 plastic cap was removed.
- Kimi-2.6: 1.
- Qwen3.5-27B: One hundred caps.
- MIMO-v2.5: Based on the video, a total of 3 plastic caps were removed.

Nature



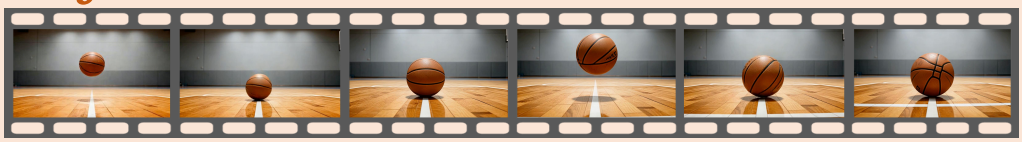
- Subcategory: Lightning
- Task Type: Temporal Reasoning
- Q: Analyzing this video, how were the trees in the video damaged?
- GT: The tree was damaged by lightning..

Model Answer:

- Seed-2.0-pro: The trees (the prominent tall tree visible in the distance) are struck by lightning (the initial bright flash is the lightning discharge): the lightning hit directly damages the tree, causing it to break apart and collapse.
- Gemini-3.1-Pro: The trees in the video were damaged by the strong winds of a tornado.
- GPT-5.4: The trees appear to have been damaged by strong wind during a storm..
- Kimi-2.6: The tall tree on the right was struck by lightning.
- Qwen3.5-27B: The trees in the video were damaged by a storm.
- MIMO-v2.5: Based on the video, the trees are being damaged by extremely strong winds, likely from a severe storm or tornado.

Figure 10: Partial model test results on **MOMENT-VIDEO** (Part III).

AIGC



- Subcategory: Artifacts
- Task Type: Temporal Occurrence
- Q: Analyzing this video, does it contain anything that violates the real world? Please state the specific reason.
- GT: In the video, the basketball loses its bounce after landing and then suddenly bounces back up. This doesn't conform to real-world phenomena.

Model Answer:

- Seed-2.0-pro: There is content that violates real-world logic, the specific reasons are as follows: 1. The surface markings of the basketball change randomly every time it bounces. 2. The basketball also does not show the slight deformation that a real inflated basketball should produce when it collides with the ground.
- Gemini-3.1-Pro: Yes, this video contains elements that violate the real world. The texture and pattern of the basketball change drastically and impossibly throughout the video. In the real world, a basketball's physical appearance remains constant while bouncing. However, in this video, as the ball bounces, the black lines on its surface continuously morph, change shape, and become highly distorted, particularly noticeable in the later frames compared to the earlier ones.
- GPT-5.4: Yes. The basketball appears to hover or rise off the floor on its own, which violates normal real-world physics.
- Kimi-2.6: Yes. The basketball violates real-world physics by spontaneously launching into the air from a resting position on the without any visible external force, person, or mechanism acting upon it.
- Qwen3.5-27B: Yes, the video contains elements that violate the real world. The basketball is shown floating in mid-air without any visible support or force acting upon it, which defies the laws of physics.
- MIMO-v2.5: No, the video does not appear to contain anything that violates the real world.

Figure 11: Partial model test results on **MOMENT-VIDEO** (Part IV).

Open-ended LLM-as-Judge Prompt

System prompt:

```
You are a strict and fair evaluator. Output valid JSON only.
```

User message:

Evaluate whether the model answer is semantically consistent with the reference answer.

Judging rules:

1. Judge meaning, not wording.
2. Accept paraphrases, synonyms, abbreviations, and equivalent naming variants.
3. Mark as consistent if the model answer fully covers the reference answer's meaning, even with extra harmless details.
4. Mark as inconsistent if it misses a key point, contradicts the reference answer, or changes important facts, including entity, action, order, quantity, identity, or existence.
5. Mark as inconsistent if the model answer is too vague to support the same meaning.
6. If uncertain, choose consistent only when a reasonable reader would conclude full semantic coverage.

Return JSON only, with no markdown and no extra text:

```
{"is_consistent": true/false, "reason": "one-sentence
explanation"}
```

Reference answer:
`{reference_answer}`

Model answer:
`{model_answer}`

Figure 12: Open-ended LLM-as-Judge prompt used to evaluate semantic consistency between the model prediction and the reference answer.