TimeBlind: A Spatio-Temporal Compositionality Benchmark for Video LLMs

Anonymous ACL submission

Abstract

Fine-grained spatio-temporal understanding is essential for video reasoning and embodied AI. Yet, while Multimodal Large Language Models (MLLMs) master static semantics, their grasp of temporal dynamics remains brittle. We present TimeBlind, a diagnostic benchmark for compositional spatio-temporal understanding. Inspired by cognitive science, TimeBlind categorizes fine-grained temporal understanding into three levels: recognizing atomic events, characterizing event properties, and reasoning about event interdependencies. Unlike benchmarks that conflate recognition with temporal reasoning, TimeBlind leverages a minimal-pairs paradigm: video pairs share identical static visual content but differ solely in temporal structure, utilizing complementary questions to neutralize language priors. Evaluating over 20 state-of-the-art MLLMs (e.g., GPT-5, Gemini 3 Pro) on 600 curated instances (2400 video-question pairs), reveals that the Instance Accuracy (correctly distinguishing both videos in a pair) of the best performing MLLM is only 48.2%, far below the human performance (98.2%). These results demonstrate that even frontier models exhibit significant deficiencies in temporal reasoning, positioning TimeBlind as a vital diagnostic tool for next-generation video understanding.

1 Introduction

Fine-grained spatio-temporal understanding is fundamental for long-horizon video reasoning (Wang et al., 2025a; Pan et al., 2025) and embodied AI (Tang et al., 2025; Yang et al., 2025b). Beyond simply recognizing *what* is present, an intelligent system must infer *what changes*, *how it changes*, and *how multiple changes compose* into causal structures. While recent Multimodal Large Language Models (MLLMs) (Pichai et al., 2025; OpenAI, 2025; Bai et al., 2025; Clark et al., 2025; Cho et al., 2025) demonstrate impressive performance on general benchmarks, their "sense of time" remains surprisingly brittle. As shown in Figure 1, even frontier models (e.g., GPT-5 and Gemini 3 Pro) struggle to distinguish atomic actions (e.g., *shaking* vs. *holding*) in videos as short as 10 seconds. Moreover, models often misjudge relative dynamics (e.g., *accelerating* vs. *decelerating*), and frequently fail to resolve linguistic temporal connectives (e.g., *"as soon as"*). This discrepancy suggests that current benchmark scores (Jang et al., 2017a; Yu et al., 2019; Li et al., 2024c; Zhou et al., 2025a; Fu et al., 2025; Hu et al., 2025) severely overestimate genuine temporal understanding capabilities.

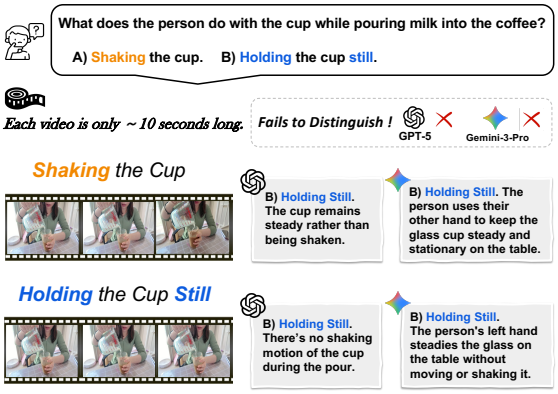

Figure 1: **An example video pair that shares identical static visual content but differs solely in motion dynamics.** The top video shows a person shaking a cup, while the bottom video shows them holding it still. Even the most advanced models like GPT-5 and Gemini 3 Pro fail to distinguish the actions in the video pair.

The root issue lies in evaluation design. Existing benchmarks (Yu et al., 2019; Li et al., 2024c; Liu et al., 2024; Fu et al., 2025) rarely *isolate* temporal structure as the sole discriminative factor. As a result, many models exploit "static shortcuts"—correlating visual entities with answers without modeling time (Lei et al., 2023; Krojer et al., 2025). Furthermore, language priors (Li et al., 2024b) allow models to guess answers based

Table 1: **Comparison with prior temporal video benchmarks.** We compare coverage of temporal reasoning categories and anti-shortcut design features. TimeBlind uniquely covers all 13 Allen temporal relations[†], includes causal and comparative reasoning, and employs both paired videos and complementary questions to eliminate static and linguistic shortcuts. The rightmost column shows the gap between human and best model performance, indicating benchmark difficulty.

		Temporal Reasoning Coverage					Anti-Shortcut Design		
Benchmark	Size	Event Types	Event Attributes	Temporal Topologies	Causal	Cross-Event Comparison	Video Pairs	Compl. Questions	Human–Model Gap
MVBench (Li et al., 2024c)	4K	2	2	2	✗	✗	✗	✗	–
TOMATO (Shangguan et al., 2024)	1.4K	2	4	2	✗	✗	✗	✗	57.3%
Vinoground (Zhang et al., 2024a)	2K	1	1	2	✗	✗	✓	✗	40.0%
TempCompass (Liu et al., 2024)	4K	2	2	2	✗	✓	✓	✗	–
TimeBlind (Ours)	2.4K	2	6	13[†]	✓	✓	✓	✓	50.0%

[†]Covers all 13 event relations in Allen's Interval Algebra (Allen, 1983): *before, after, meets, met-by, overlaps, overlapped-by, starts, started-by, finishes, finished-by, during, contains, equals*.

on textual plausibility. To truly *diagnose* understanding, an evaluation must hold the visual content constant and vary only the temporal dynamics.

To address this, we introduce **TimeBlind**, a diagnostic benchmark for compositional spatio-temporal understanding (Table 1). TimeBlind adopts a minimal-pairs design: each instance contains two videos with near-identical static visual content that differ *solely* in their temporal structure. To eliminate language priors, we utilize complementary questions where the correct answer flips between the paired videos. By removing static and linguistic shortcuts, TimeBlind forces models to rely exclusively on temporal evidence. Unlike large-scale noisy benchmarks, TimeBlind prioritizes high-fidelity diagnostic precision. Similar to Winoground (Thrush et al., 2022), each instance serves as a rigorous test for a specific cognitive primitive, where high-quality annotations are prioritized over scale.

Crucially, unlike prior work (Zhang et al., 2024a; Krojer et al., 2025), TimeBlind evaluates logical composition, not just perception. Inspired by cognitive science, we extend the theory of *Image Compositionality* (Krishna et al., 2017; Li et al., 2024a) to the temporal domain, organizing the benchmark around a hierarchical taxonomy: (i) *Events*, addressing what happened (e.g., changes in object attributes or action understanding); (ii) *Event Attributes*, describing how events unfold (e.g., speed, magnitude of change); and (iii) *Structural Event Logic*, examining how multiple events are composed (e.g., temporal topology, causality, and cross-event comparison). Within this framework, we encompass a diverse set of 11 fine-grained categories in real-world scenarios. For example, regarding *Temporal Topology* in *Structural Event Logic*, we incorporate Allen's Interval Algebra (Allen, 1983) by constructing videos that cover all 13 temporal event relations (e.g., *overlaps*, *meets*, *equals*) and designing questions that distinguish specific topology differences within each pair, going beyond simple sequencing relations (e.g., *before*, *after*) used in prior work.

We evaluate over 20 state-of-the-art MLLMs, including leading proprietary models (e.g., Gemini 3 Pro (Pichai et al., 2025), GPT-5 (OpenAI, 2025)) as well as strong open-source models (e.g., Qwen3-VL (Bai et al., 2025), Molmo2 (Clark et al., 2025), PLM (Cho et al., 2025)), on 600 curated instances (2,400 video-question pairs). We find that even the top-performing model, Gemini 3 Pro, achieves only 48.2% Instance Accuracy (I-Acc), which requires correctly distinguishing both videos in a pair, far below human performance (98.2%). Furthermore, while advanced models like Gemini 3 Pro and GPT-5 demonstrate reasonable proficiency in recognizing isolated events (achieving 49.2% and 58.3% I-Acc, respectively), their performance degrades sharply on event attributes—such as Speed (slowly vs. rapidly), and Force (forcefully vs. gently)—dropping to 36.7% and 32.3% I-Acc. Moreover, extensive ablation studies demonstrate that model performance remains poor even with increased input frames or test-time reasoning, as GPT-5 gains only 3.3% I-Acc, indicating models remain "time-blind" and fragile in temporal reasoning. We believe TimeBlind will serve as a vital diagnostic tool for evaluating and developing MLLMs capable of genuine temporal logic.

In summary, our main contributions are:

- **TimeBlind Benchmark.** A diagnostic, minimal-pairs benchmark that isolates temporal structure while minimizing static shortcuts and language

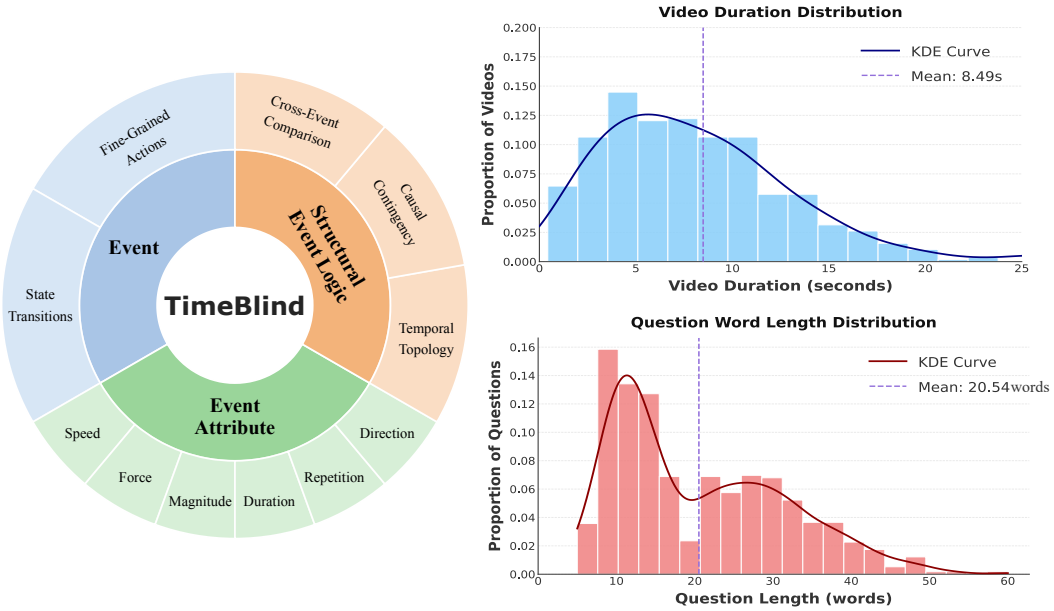

Figure 2: **TimeBlind Taxonomy and Statistics. Left:** We structure the evaluation into 11 fine-grained spatio-temporal compositional categories spanning three high-level aspects: Atomic Events (*what changes*), Parametric Event Attributes (*how it changes*), and Structural Event Logic (*how events compose*). **Top Right:** Distribution of video lengths across the benchmark, showing that most videos fall within the 0–15 seconds. **Bottom Right:** Distribution of question word counts, indicating that most questions are under 30 words. Overall, our benchmark features a structured taxonomy with diverse categories while maintaining short videos and concise questions.

priors.

- **Taxonomy of Temporal Compositionality.** A cognitive hierarchy—*Events*, *Attributes*, and *Structural Logic*—that guides systematic benchmark construction and evaluation.
- **Diagnostic Findings.** An evaluation of over 20 SOTA MLLMs revealing a significant gap between perceived and actual temporal reasoning capabilities.

2 Related Work

VideoQA Benchmarks. Early VideoQA datasets (Yu et al., 2019; Jang et al., 2017b; Patraucean et al., 2023; Xiao et al., 2021; Lei et al., 2018; Xu et al., 2016; Wu et al., 2024a) focus on simple scenarios with short clips and limited question types. Recent benchmarks address comprehensive evaluation (Fu et al., 2025; Li et al., 2024c; Ma et al., 2025), complex reasoning (Nagrani et al., 2025; Cheng et al., 2025; Hu et al., 2025; Song et al., 2025), long-form understanding (Mangalam et al., 2023; Wang et al., 2025a; Wu et al., 2024b; Song et al., 2024; Zhou et al., 2025a; Yang et al., 2025a), and domain-specific settings (Pan et al., 2025; Long et al., 2025; Yi et al., 2025). However, most do not isolate temporal structure as the sole discriminative factor, allowing models to exploit static shortcuts—relying on object co-occurrence or language priors without genuinely modeling temporal dynamics (Lei et al., 2023; Krojer et al., 2025; Li et al., 2024b).

Evaluating Spatio-Temporal Compositionality. To assess the spatio-temporal understanding capabilities of MLLMs, several recent benchmarks have introduced diverse aspects of visual reasoning. In the image understanding domain, BLINK (Fu et al., 2024) reformats and groups classic vision problems into perception-centric multiple-choice questions and carefully removes language priors. In the video understanding domain, several works explicitly target temporal understanding by designing temporally challenging questions (Johnson et al., 2017; Shangguan et al., 2024; Cai et al., 2024; Xue et al., 2025). Beyond single-video evaluation, a growing line of work adopts paired video–question protocols to more precisely diagnose temporal reasoning. TempCompass (Liu et al., 2024) constructs paired videos by systematically manipulating an original video, such as reversing playback or altering temporal speed, thereby isolating temporal understanding from "static shortcuts." VinoGround (Zhang et al., 2024a) further strengthens this paradigm by requiring models to answer identical questions over

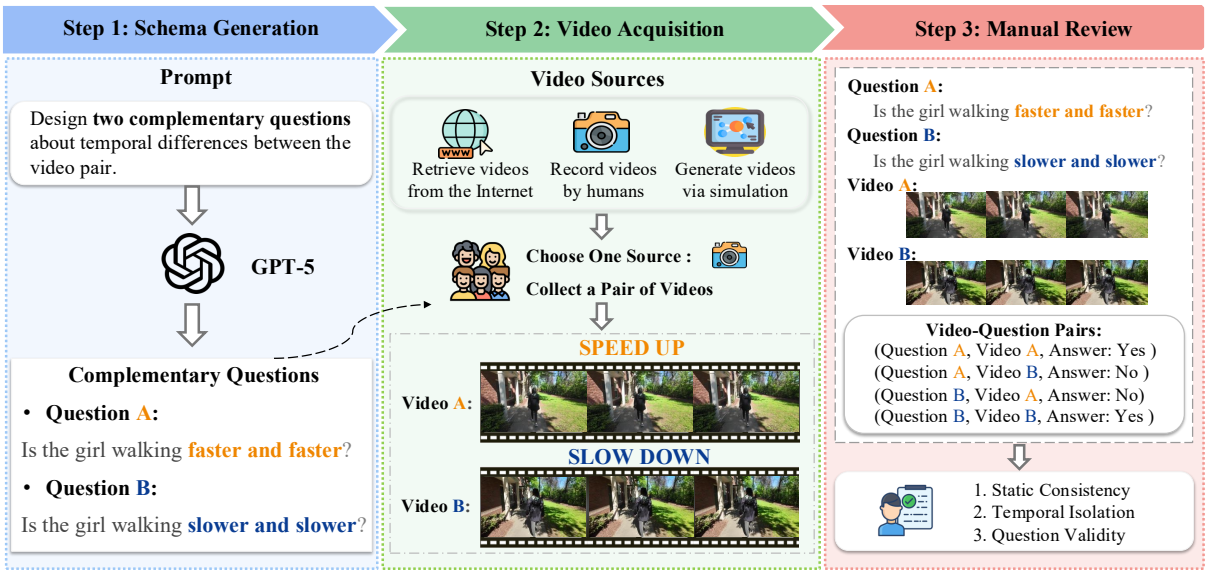

Figure 3: **Overview of the TimeBlind data construction pipeline. Stage 1 (Schema Generation):** We prompt GPT-5 to generate paired complementary questions targeting temporal differences. **Stage 2 (Video Acquisition):** We collect one video pair that matches the generated schema from one of the following sources: (i) Retrieving videos from the internet, (ii) Recording videos with humans, or (iii) Generating videos via simulation (e.g., Unity). We then pair these videos with the questions to form a candidate TimeBlind instance. **Stage 3 (Manual Review):** Human annotators manually review each instance to ensure: (i) *Static Consistency* (videos share identical static content), (ii) *Temporal Minimality* (the pair differs only in the targeted temporal factor), and (iii) *Question Validity* (QA pairs are clear and correct).

paired videos, where the correct answer is determined solely by temporal differences, effectively removing language priors. Follow-up works like GLIMPSE (Zhou et al., 2025b) and MVP (Krojer et al., 2025) scale this approach to test physical and visual-centric reasoning. Different from prior works, TIMEBLIND achieves spatio-temporal compositionality through a carefully curated formal taxonomy. Drawing inspiration from cognitive event perception (Bach, 1986) and extending the theory of image compositionality (Krishna et al., 2017; Thrush et al., 2022; Li et al., 2024b), we decompose temporal reasoning into atomic *Events*, parametric *Event Attributes*, and *Structural Logic*.

3 The TimeBlind Benchmark

In this section, we describe the design of TimeBlind, a diagnostic benchmark for rigorously evaluating compositional spatio-temporal understanding. Our design is grounded in two principles: (i) a temporal minimal-pair protocol that minimizes visual shortcuts and language priors, and (ii) a cognition-inspired compositionality taxonomy that decomposes temporal reasoning into three primitives: *Events*, *Event Attributes*, and *Structural Event Logic*.

3.1 Task Formulation

We structure each instance as $\mathcal{T} = (v_1, v_2, q_1, q_2)$, following the design below. This design creates a discriminative challenge where shortcut solutions are effectively "canceled out," forcing the model to rely on temporal evidence.

Temporal Minimal Pairs. The video pair (v_1, v_2) constitutes a minimal pair: the videos share identical static content (e.g., objects, background) but differ along an isolated temporal axis defined in our taxonomy (Section 3.2). Consequently, a model cannot distinguish v_1 from v_2 without explicitly modeling the temporal features (e.g., *opening* vs. *closing*).

Logical Complementarity. The questions (q_1, q_2) are complementary: for any question, the ground-truth answer flips between the two videos (e.g., $Ans(v_1, q_j) \neq Ans(v_2, q_j)$ for $j \in \{1, 2\}$). This design neutralizes language priors, preventing models from exploiting textual plausibility.

Diagnostic Metrics. Following rigorous diagnostic protocols (Thrush et al., 2022; Li et al., 2024b), we report a hierarchy of metrics to mitigate shortcuts and assess the models' genuine temporal understanding:

- **Accuracy (Acc):** Standard accuracy computed

over all individual video–question trials.

- **Video Accuracy (V-Acc):** Measures visual consistency; correct only if the model answers *both* questions correctly for a single video v_j.
- **Question Accuracy (Q-Acc):** Measures textual consistency; correct only if the model answers question q_i correctly for *both* videos.
- **Instance Accuracy (I-Acc):** An instance is correct *if and only if* the model solves all four trials. I-Acc is our primary proxy for spatio-temporal understanding, as it necessitates reliably distinguishing the temporal difference between the paired videos.

3.2 A Hierarchy of Temporal Composition

To systematically evaluate temporal understanding, we organize TimeBlind around a compositional taxonomy mirroring a bottom-up cognitive process: from recognizing atomic changes (*Events*), to measuring their intrinsic properties (*Event Attributes*), and finally reasoning about their interdependencies (*Structural Event Logic*), shown in Figure 2.

I. Events. This foundational primitive tests the detection of atomic visual changes, demanding that models distinguish actual temporal evolution from static object presence.

- **Fine-Grained Actions:** Discriminating actions sharing visual context but differing in temporal dynamics (e.g., *opening* vs. *closing*) or semantic nuance (e.g., *placing* vs. *pushing*).
- **State Transitions:** Tracking object property changes over time (shape, size, color, emotion)—e.g., liquid changing from clear to opaque.

II. Event Attributes. This primitive tests the perception of *how* an event unfolds, requiring sensitivity to continuous or qualitative parameters rather than categorical recognition.

- **Kinematics:** Properties governing motion in space-time, including *Speed* (fast/slow), *Direction* (forward/backward), *Duration* (brief/long), and *Repetition* (once/twice).
- **Dynamics & Manner:** Properties reflecting physical forces or execution style, including *Force* (gentle/forceful) and *Magnitude* (slight/broad movement).

III. Structural Event Logic. This level tests *how* multiple events relate and compose into higher-order structures.

- **Temporal Topology:** We adopt Allen's Interval Algebra (Allen, 1983) to comprehensively evaluate interval relations. Unlike prior work limited to simple sequencing (*before/after*), we cover all 13 relations in our videos, such as *overlaps*, *meets*, *starts*, and *during*.
- **Causal Contingency:** Identifying causal links between events (e.g., identifying which event *caused* the effect).
- **Cross-Event Comparison:** Comparative reasoning across distinct events (e.g., "Does the person hold the cup *longer* than the book?").

3.3 Data Construction Pipeline

We use a three-stage pipeline to collect high-quality TimeBlind data, as shown in Figure 3.

Step 1: Schema Generation. We prompt frontier LLMs (e.g., GPT-5) with the taxonomy definitions in Section 3.2 to generate structured specifications. Each output includes a pair of video descriptions (e.g., "*a girl walks in a park and gradually speeds up*" and "*a girl walks in a park and gradually slows down*") that differ only in the target temporal factor (e.g., "*Speed*"), along with complementary questions (e.g., "*Is the girl walking faster and faster?*" and "*Is the girl walking slower and slower?*").

Step 2: Video Acquisition. We collect video pairs from three sources: (i) internet retrieval (extracting temporally distinct segments from the same source video), (ii) human recording, or (iii) simulation (e.g., Unity) for precise temporal control. We then pair the collected videos with the questions to form a candidate TimeBlind instance.

Step 3: Rigorous Human Verification. We recruited 8 annotators for quality control. Each instance was independently reviewed by two annotators and retained only if all constituent video-question pairs received unanimous approval. Annotators assessed three criteria: (i) *Static Consistency*, ensuring that paired videos share identical static content; (ii) *Temporal Minimality*, verifying that they differ only in the targeted temporal factor; and (iii) *Question Validity*, confirming that the QA pairs are clear and correct. Overall, the inter-annotator agreement reached 94.3%, with a Cohen's κ of 0.835. The review took 62.2 hours in total. Per-category agreement scores are reported in Appendix A.2.

Statistics. TimeBlind comprises 2,400 curated video-question pairs, consisting of binary and two-choice multiple-choice questions. The video sources are distributed as follows: 24.0% from Internet Retrieval, 57.7% from Human Recording, and 18.3% from Simulation. Regarding category distribution, we maintain an approximate 1:1:1 ra-

Table 2: **Main Results on TimeBlind.** We use uniform sampling at 1 FPS and evaluate all models using default configurations. Metrics are reported following Section 3, with I-Acc as the primary metric. The table is divided into *Open-Source (grouped by size: $< 10B$ and $> 10B$)*, and *Closed-Source models*. Our results suggest that all models perform poorly on fine-grained temporal video understanding, with none of the methods achieving over 50% I-Acc. Best results are shown in **bold**.

Model	Size	Q-Acc (%)	V-Acc (%)	Acc (%)	I-Acc (%)
Random Chance	-	25.0	25.0	50.0	6.3
Human Evaluation	-	98.8	98.9	99.3	98.2
Open-Source Models					
Model Size $\leq 10B$					
Keye-VL-1.5 (Team et al., 2025)	8B	17.5	26.5	55.6	6.6
LLaVA-Video (Zhang et al., 2024b)	7B	18.1	34.8	57.8	9.6
InternVL3.5 (Wang et al., 2025b)	8B	30.9	39.0	59.3	13.3
PLM-8B (Cho et al., 2025)	8B	24.4	32.3	60.6	13.9
F-16 (Li et al., 2025)	7B	23.5	39.7	60.5	14.5
video-SALMONN 2+ (Tang et al., 2025)	7B	26.3	41.7	61.7	16.2
MiniCPM-V-4.5 (Yu et al., 2025)	8B	31.7	43.6	62.7	18.3
VideoChat-Flash (Li et al., 2024d)	7B	31.7	45.4	64.9	19.5
Qwen3-VL (Bai et al., 2025)	8B	34.5	43.6	64.8	19.7
GLM-4.1V (Hong et al., 2025)	9B	32.3	45.2	62.0	19.7
Eagle2.5 (Chen et al., 2025)	8B	32.2	45.7	64.7	20.2
Molmo2 (Clark et al., 2025)	8B	**41.0**	**52.4**	**68.7**	**31.2**
Model Size $\geq 30B$					
InternVL 3.5 (Wang et al., 2025b)	30B	31.4	41.3	61.1	17.0
LLaVA-Video (Zhang et al., 2024b)	72B	28.8	45.0	63.2	18.5
Qwen3-VL (Bai et al., 2025)	235B	**38.1**	**51.7**	**66.9**	**25.8**
Closed-Source Models					
Claude Sonnet 4.5 (Anthropic, 2025)	unknown	30.6	38.5	59.2	13.6
Qwen3-VL Plus (Bai et al., 2025)	unknown	38.7	52.6	67.3	26.0
GPT-5 mini (OpenAI, 2025)	unknown	42.3	53.7	68.3	30.0
Gemini 2.5 Flash (Kavukcuoglu, 2025)	unknown	47.6	53.7	68.9	33.2
Gemini 2.5 Pro (Kavukcuoglu, 2025)	unknown	56.0	62.6	74.2	43.5
GPT-5 (OpenAI, 2025)	unknown	58.5	**67.2**	**77.3**	46.3
Gemini 3 Pro (Pichai et al., 2025)	unknown	**60.2**	66.0	76.2	**48.2**

tio across *Event*, *Event Attribute*, and *Structural Event Logic*. Furthermore, within the *Temporal Topology* sub-category, we ensure a balanced distribution across the 13 distinct Allen interval types.

4 Experimental Setup

Baseline Models. To thoroughly assess the challenges posed by TimeBlind, we evaluate over 20 MLLMs, including both frontier proprietary models (e.g., GPT-5 (OpenAI, 2025), Gemini 3 Pro (Pichai et al., 2025), Claude (Anthropic, 2025), and Qwen3-VL Plus (Bai et al., 2025)) and state-of-the-art open-source models (e.g., Qwen3-VL (Bai et al., 2025), InternVL 3.5 (Wang et al., 2025b), Molmo2 (Clark et al., 2025), PLM (Cho et al., 2025), Keye-VL (Team et al., 2025), MiniCPM-V-4.5 (Yu et al., 2025), GLM-4.1V (Hong et al., 2025), and Eagle 2.5 (Chen et al., 2025)).

Evaluation Metrics. We adopt the hierarchy of metrics defined in Section 3.1, including Standard Accuracy (Acc), Video Accuracy (V-Acc), Question Accuracy (Q-Acc), and our primary metric, Instance Accuracy (I-Acc).

Implementation Details. We run most open-source model experiments on 4 NVIDIA H100 GPUs, while accessing proprietary models (e.g., GPT-5, Gemini 3 Pro) and large-scale open-source models (e.g., Qwen3-VL-235B) through their official APIs. Unless otherwise noted, we sample videos at 1 FPS (the default for most models) and evaluate in a zero-shot setting.

5 Experimental Results

5.1 Main Results

In Table 2, we present the overall model performances on our benchmark, revealing that all models perform poorly in fine-grained temporal video understanding. While many models achieve relatively high standard accuracy (Acc), they struggle

to achieve good Instance Accuracy (I-Acc), which requires correctly distinguishing both videos in a pair. For instance, despite leading models like GPT-5 and Gemini 3 Pro achieving high Acc scores of 77.3% and 76.2% respectively, they only reach 46.3% and 48.2% on I-Acc. This discrepancy indicates that models remain weak in temporal understanding and suggests that high performance on Acc is often driven by shortcuts, rather than a true understanding of temporal dynamics. Another interesting finding is that Question Accuracy (Q-Acc) is consistently lower than Video Accuracy (V-Acc) across all evaluated models. This indicates that models are more prone to hallucinating answers based on textual patterns than misinterpreting visual cues. From the results, we also observe that among open-source models, Molmo2-8B—despite its smaller size—outperforms a range of sub-10B models and even surpasses the much larger Qwen3-VL-235B by 5.4% on I-Acc, becoming the leading open-source model. This suggests that with effective model design, data curation, and training process, even smaller models can capture fine-grained temporal dynamics. Lastly, our results show a significant gap between open-source and proprietary models in understanding fine-grained temporal dynamics. Specifically, Molmo2-8B still lags substantially behind GPT-5 and Gemini 3 Pro, with gaps of 15.1% and 17.0% I-Acc, respectively.

Human Evaluation. Four independent annotators evaluated the full TimeBlind dataset under a balanced workload, with each seeing only one (v, q) pair at a time and the four pairs from each instance distributed across different annotators (Table 2). Humans reached 98.2% I-Acc—50 percentage points above Gemini 3 Pro—showing that TimeBlind's temporal dynamics are clear to humans yet challenging for MLLMs.

5.2 Category-Wise Diagnosis

We analyze performance across the three high-level TimeBlind categories in Table 3. First, we observe that models consistently perform better on discrete *Events* than on *Event Attributes* or *Structural Event Logic*. For instance, GPT-5 reaches 58.3% on *Events* but drops to 32.3% on *Event Attributes*. Furthermore, performance is lowest in *Event Attributes*, which probes continuous parameters such as speed, force, and magnitude. Even the strongest proprietary models GPT-5 and Gemini 3 Pro only achieve 32.3% and 36.7%, exposing a systematic deficiency in low-level, physics-related temporal understanding. Finally, a striking disparity emerges on *Structural Event Logic* between proprietary and open-source models. While Gemini 3 Pro reaches 57.8%, the leading open-source model Qwen3-VL-235B only achieves 18.8%. These results indicate that open-source MLLMs can detect isolated events but lack the logical abilities to reason about precise relationships between multiple events, particularly regarding causality.

Table 3: **Category-Wise I-Acc (%) on TimeBlind.** We report results across the three high-level categories: *Event*, *Attribute* (Event Attribute), and *Logic* (Structural Event Logic). Models generally perform well on *Events* but struggle with *Attributes*, which require low-level physical understanding. Best results are **bolded**.

Models	Event	Attribute	Logic
Random Chance	6.3	6.3	6.3
Open-Source Models			
InternVL3.5-38B	32.4	9.4	9.4
LLaVA-Video-72B	33.3	7.3	12.0
Qwen3-VL-235B	43.6	12.5	**18.8**
Molmo2-8B	**52.0**	**20.3**	17.7
Closed-Source Models			
GPT-5	**58.3**	32.3	48.4
Gemini 3 Pro	49.2	**36.7**	**57.8**

Table 4: **Shortcut Analysis.** We report I-Acc for three shortcut baselines using GPT-5 with 1 FPS sampling: Single Frame Bias (a question paired with a randomly selected video frame), Language Only (a question without visual input), and Visual-Cue (a question with shuffled video frames). The results demonstrate that solving TimeBlind requires genuine temporal understanding.

Setting	Acc	Q-Acc	V-Acc	I-Acc
Random Chance	50.0	25.0	25.0	6.3
GPT-5 (*baseline*)	77.3	58.5	67.2	46.3
Single Frame	52.2	14.8	32.3	4.5
Language Only	47.3	9.4	26.6	1.5
Visual-Cue	49.6	10.7	23.2	3.0

5.3 Shortcut Analysis

To verify that TimeBlind requires genuine temporal understanding, we report three tests in Table 4.

- **Single-Frame Bias.** This experiment evaluates whether TimeBlind requires reasoning over a sequence rather than exploiting information from a single static frame. In this setting, the model is provided with the question and only one randomly sampled frame. The results show that GPT-5 performs poorly (4.5% I-Acc), indicating

Table 5: **Effect of Model Size.** We report the I-Acc performance of Qwen3-VL across three model sizes. The results indicate that simply scaling up model size fails to significantly improve performance.

Model Size	8B	30B	235B
I-Acc	19.7	22.5	**25.8**

that TimeBlind requires sequential modeling.

- **Language-Only Bias.** This experiment evaluates the influence of language priors within TimeBlind. In this setting, the model is provided with only the question without any visual information. As shown in Table 4, GPT-5 achieves only 1.5% I-Acc, demonstrating that visual information is essential for our benchmark setting.
- **Visual-Cue Shortcuts.** Finally, we examine whether the benchmark can be solved by exploiting static visual cues that happen to correlate with temporal dynamics. In this setting, we shuffle the order of the input frames sampled at 1 FPS to ensure that successfully completing the task requires an understanding of the temporal sequence rather than merely detecting specific objects. GPT-5 achieves only 3.0% I-Acc in this setting, indicating that solving TimeBlind requires strict temporal understanding.

Moreover, Acc scores across all three shortcut settings hover around random chance, suggesting that correctly answering even a single question in TimeBlind requires strict temporal understanding.

5.4 Additional Analysis

We conduct ablation studies on three factors: model size, input FPS, and inference-time reasoning. More ablations can be found in Appendix A.7.

Effect of Model Size. Table 5 shows that increasing model size brings only limited performance gains, with Qwen3-VL improving by merely 6.1% in I-Acc when scaled from 8B to 235B. Simply scaling model size does not yield robust spatio-temporal understanding.

Effect of FPS. Table 6 shows that increasing the sampling rate yields only modest gains. All evaluated models peak at 5 FPS, with Gemini 3 Pro reaching the best score (56.2% I-Acc)—still 42.0% below human performance. Performance drops at 10 FPS across models, suggesting that denser frame sampling does not reliably improve temporal understanding and may even hinder reasoning due to longer input sequences.

Test-Time Scaling (Reasoning). Table 7 shows inference-time reasoning results. Due to computational cost, we evaluate on a random 30% subset of the data. For Qwen3-VL, *Thinking* variants outperform *Instruct* models, with the 235B model gaining 10.4% but still achieving only 36.3% I-Acc. For GPT-5, *Thinking* yields modest gains, achieving 49.6% I-Acc—still far below human performance (98.2%). While reasoning helps, it remains insufficient to solve TimeBlind.

Table 6: **Effect of Input FPS.** We report I-Acc across different input frame rates. The best score—Gemini 3 Pro at 5 FPS (56.2%)—still trails human performance by 42.0%. All models peak at 5 FPS and degrade at higher FPS, indicating that higher FPS does not consistently improve temporal understanding.

Model	1 FPS	5 FPS	10 FPS
Human	98.2	98.2	98.2
Qwen3-VL-235B	25.8	35.0	26.5
Molmo2-8B	31.2	31.7	29.2
GPT-5	46.3	51.0	49.3
Gemini 3 Pro	48.2	56.2	49.6

Table 7: **Inference-Time Reasoning.** We compare standard Qwen3-VL and GPT-5 models against their reasoning-enabled *Thinking* counterparts across various settings, using the I-Acc metric at 1 FPS sampling. Results show that sufficient inference-time reasoning depth improves temporal understanding, but is far from sufficient to solve TimeBlind.

Model	**Mode**	**I-Acc (%)**	**Δ**
Qwen3-VL-8B	Standard	19.6	-
	Thinking	27.8	+8.2
Qwen3-VL-235B	Standard	25.8	-
	Thinking	36.3	**+10.4**
GPT-5	Standard	46.3	-
	Thinking	**49.6**	+3.3

6 Conclusion

We introduce TimeBlind, a diagnostic benchmark for compositional spatio-temporal reasoning in MLLMs. Built on a cognitive taxonomy of *events*, *event attributes*, and *structural event logic*, together with a strict minimal-pair design, TimeBlind isolates temporal understanding from static and linguistic shortcuts. Evaluation results show that current state-of-the-art models remain largely "time-blind", falling far behind human performance. We hope TimeBlind will guide future work toward genuinely time-aware MLLMs.

7 Limitations

While TimeBlind offers a rigorous diagnostic for compositional spatio-temporal understanding, we acknowledge several limitations. First, TimeBlind focuses on short-form videos (average 8.49 seconds), which isolates fine-grained temporal dynamics but leaves long-horizon temporal reasoning over minutes or hours outside its scope. Second, all questions are written in English, leaving the assessment of multilingual temporal understanding as an open direction. Third, we adopt binary and two-choice multiple-choice formats to enable unambiguous scoring under our minimal-pair protocol; evaluating temporal understanding through open-ended generation remains a complementary direction. Finally, certain *Event Attribute* sub-categories (e.g., Force, Magnitude) inherently involve qualitative judgments whose boundaries can be subjective; while our annotation pipeline maintains substantial inter-annotator agreement ($\kappa = 0.84$ for *Event Attributes*), residual ambiguity is intrinsic to these concepts. We view these as natural directions for extending TimeBlind, rather than as challenges to its diagnostic conclusions.

References

James F Allen. 1983. Maintaining knowledge about temporal intervals. *Communications of the ACM*, 26(11):832–843.

Anthropic. 2025. Introducing claude sonnet 4.5. Anthropic News.

Emmon Bach. 1986. The algebra of events. *Linguistics and philosophy*, pages 5–16.

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

Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, and 1 others. 2024. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*.

Guo Chen, Zhiqi Li, Shihao Wang, Jindong Jiang, Yicheng Liu, Lidong Lu, De-An Huang, Wonmin Byeon, Matthieu Le, Max Ehrlich, Tong Lu, Limin Wang, Bryan Catanzaro, Jan Kautz, Andrew Tao, Zhiding Yu, and Guilin Liu. 2025. Eagle 2.5: Boosting long-context post-training for frontier vision-language models.

Junhao Cheng, Yuying Ge, Teng Wang, Yixiao Ge, Jing Liao, and Ying Shan. 2025. Video-holmes: Can mllm think like holmes for complex video reasoning? *arXiv preprint arXiv:2505.21374*.

Jang Hyun Cho, Andrea Madotto, Effrosyni Mavroudi, Triantafyllos Afouras, Tushar Nagarajan, Muhammad Maaz, Yale Song, Tengyu Ma, Shuming Hu, Suyog Jain, and 1 others. 2025. Perceptionlm: Open-access data and models for detailed visual understanding. *arXiv preprint arXiv:2504.13180*.

Christopher Clark, Jieyu Zhang, Zixian Ma, Jae Sung Park, Mohammadreza Salehi, Rohun Tripathi, Sangho Lee, Zhongzheng Ren, Chris Dongjoo Kim, Yinuo Yang, Vincent Shao, Yue Yang, Weikai Huang, Ziqi Gao, Taira Anderson, Jianrui Zhang, Jitesh Jain, George Stoica, Winson Han, and 2 others. 2025. Molmo2: Open weights and data for vision-language models with video understanding and grounding. Technical report, Allen Institute for AI.

Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, and 1 others. 2025. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24108–24118.

Xingyu Fu, Yushi Hu, Bangzheng Li, Yu Feng, Haoyu Wang, Xudong Lin, Dan Roth, Noah A Smith, Wei-Chiu Ma, and Ranjay Krishna. 2024. Blink: Multimodal large language models can see but not perceive. In *European Conference on Computer Vision*, pages 148–166. Springer.

Wenyi Hong, Wenmeng Yu, Xiaotao Gu, Guo Wang, Guobing Gan, Haomiao Tang, Jiale Cheng, Ji Qi, Junhui Ji, Lihang Pan, and 1 others. 2025. Glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. *arXiv preprint arXiv:2507.01006*.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. 2025. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017a. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017b. Tgif-qa: Toward spatio-temporal reasoning in visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2758–2766.

Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2901–2910.

Koray Kavukcuoglu. 2025. Gemini 2.5: Our most intelligent ai model. Google Keyword Blog. Last updated March 26, 2025.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, and 1 others. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123(1):32–73.

Benno Krojer, Mojtaba Komeili, Candace Ross, Quentin Garrido, Koustuv Sinha, Nicolas Ballas, and Mahmoud Assran. 2025. A shortcut-aware video-qa benchmark for physical understanding via minimal video pairs. *arXiv preprint arXiv:2506.09987*.

Jie Lei, Tamara Berg, and Mohit Bansal. 2023. Revealing single frame bias for video-and-language learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 487–507.

Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. 2018. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 conference on empirical methods in natural language processing*, pages 1369–1379.

Baiqi Li, Zhiqiu Lin, Deepak Pathak, Jiayao Li, Yixin Fei, Kewen Wu, Tiffany Ling, Xide Xia, Pengchuan Zhang, Graham Neubig, and 1 others. 2024a. Genai-bench: Evaluating and improving compositional text-to-visual generation. *arXiv preprint arXiv:2406.13743*.

Baiqi Li, Zhiqiu Lin, Wenxuan Peng, Jean de Dieu Nyandwi, Daniel Jiang, Zixian Ma, Simran Khanuja, Ranjay Krishna, Graham Neubig, and Deva Ramanan. 2024b. Naturalbench: Evaluating vision-language models on natural adversarial samples. *Advances in Neural Information Processing Systems*, 37:17044–17068.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, and 1 others. 2024c. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206.

Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, and 1 others. 2024d. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*.

Yixuan Li, Changli Tang, Jimin Zhuang, Yudong Yang, Guangzhi Sun, Wei Li, Zejun Ma, and Chao Zhang. 2025. Improving llm video understanding with 16 frames per second. *arXiv preprint arXiv:2503.13956*.

Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. 2024. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*.

Xinwei Long, Kai Tian, Peng Xu, Guoli Jia, Jingxuan Li, Sa Yang, Yihua Shao, Kaiyan Zhang, Che Jiang, Hao Xu, Yang Liu, Jiaheng Ma, and Bowen Zhou. 2025. Adsqa: Towards advertisement video understanding. In *ICCV*.

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

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

Arsha Nagrani, Sachit Menon, Ahmet Iscen, Shyamal Buch, Ramin Mehran, Nilpa Jha, Anja Hauth, Yukun Zhu, Carl Vondrick, Mikhail Sirotenko, and 1 others. 2025. Minerva: Evaluating complex video reasoning. *arXiv preprint arXiv:2505.00681*.

OpenAI. 2025. GPT-5 system card. Accessed: 2026-01-05.

Yulu Pan, Ce Zhang, and Gedas Bertasius. 2025. Basket: A large-scale video dataset for fine-grained skill estimation. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 28952–28962.

Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, and 1 others. 2023. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761.

Sundar Pichai, Demis Hassabis, and Koray Kavukcuoglu. 2025. A new era of intelligence with gemini 3. The Keyword (Google Blog). Accessed: 2026-01-05.

Ziyao Shangguan, Chuhan Li, Yuxuan Ding, Yanan Zheng, Yilun Zhao, Tesca Fitzgerald, and Arman Cohan. 2024. Tomato: Assessing visual temporal reasoning capabilities in multimodal foundation models. *arXiv preprint arXiv:2410.23266*.

Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, and 1 others. 2024. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232.

Enxin Song, Wenhao Chai, Weili Xu, Jianwen Xie, Yuxuan Liu, and Gaoang Wang. 2025. Video-mmlu: A massive multi-discipline lecture understanding benchmark. *arXiv preprint arXiv:2504.14693*.

Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, and 1 others. 2025. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*.

Kwai Keye Team, Biao Yang, Bin Wen, Changyi Liu, Chenglong Chu, Chengru Song, Chongling Rao, Chuan Yi, Da Li, Dunju Zang, and 1 others. 2025. Kwai keye-vl technical report. *arXiv preprint arXiv:2507.01949*.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. 2022. Winoground: Probing vision and language models for visio-linguistic compositionality. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5238–5248.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Ming Ding, Xiaotao Gu, Shiyu Huang, Bin Xu, and 1 others. 2025a. Lvbench: An extreme long video understanding benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 22958–22967.

Weiyun Wang, Zhangwei Gao, Lixin Gu, Hengjun Pu, Long Cui, Xingguang Wei, Zhaoyang Liu, Linglin Jing, Shenglong Ye, Jie Shao, and 1 others. 2025b. Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. *arXiv preprint arXiv:2508.18265*.

Bo Wu, Shoubin Yu, Zhenfang Chen, Joshua B Tenenbaum, and Chuang Gan. 2024a. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*.

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

Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 2021. Next-qa: Next phase of question-answering to explaining temporal actions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9777–9786.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. 2016. Msr-vtt: A large video description dataset for bridging video and language. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 5288–5296.

Zihui Xue, Mi Luo, and Kristen Grauman. 2025. Seeing the arrow of time in large multimodal models. *arXiv preprint arXiv:2506.03340*.

Jingkang Yang, Shuai Liu, Hongming Guo, Yuhao Dong, Xiamengwei Zhang, Sicheng Zhang, Pengyun Wang, Zitang Zhou, Binzhu Xie, Ziyue Wang, and 1 others. 2025a. Egolife: Towards egocentric life assistant. *arXiv preprint arXiv:2503.03803*.

Shusheng Yang, Jihan Yang, Pinzhi Huang, Ellis Brown, Zihao Yang, Yue Yu, Shengbang Tong, Zihan Zheng, Yifan Xu, Muhan Wang, and 1 others. 2025b. Cambrian-s: Towards spatial supersensing in video. *arXiv preprint arXiv:2511.04670*.

Han Yi, Yulu Pan, Feihong He, Xinyu Liu, Benjamin Zhang, Oluwatumininu Oguntola, and Gedas Bertasius. 2025. Exact: A video-language benchmark for expert action analysis. *arXiv preprint arXiv:2506.06277*.

Tianyu Yu, Zefan Wang, Chongyi Wang, Fuwei Huang, Wenshuo Ma, Zhihui He, Tianchi Cai, Weize Chen, Yuxiang Huang, Yuanqian Zhao, and 1 others. 2025. Minicpm-v 4.5: Cooking efficient mllms via architecture, data, and training recipe. *arXiv preprint arXiv:2509.18154*.

Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 2019. Activitynet-qa: A dataset for understanding complex web videos via question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 9127–9134.

Jianrui Zhang, Mu Cai, and Yong Jae Lee. 2024a. Vinoground: Scrutinizing lmms over dense temporal reasoning with short videos. *arXiv preprint arXiv:2410.02763*.

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

Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Zhengyang Liang, Shitao Xiao, Minghao Qin, Xi Yang, Yongping Xiong, Bo Zhang, and 1 others. 2025a. Mlvu: Benchmarking multi-task long video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 13691–13701.

Yiyang Zhou, Linjie Li, Shi Qiu, Zhengyuan Yang, Yuyang Zhao, Siwei Han, Yangfan He, Kangqi Li, Haonian Ji, Zihao Zhao, and 1 others. 2025b. Glimpse: Do large vision-language models truly think with videos or just glimpse at them? In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 27830–27844.

A Appendix

This document supplements the main paper by providing additional details on the TimeBlind benchmark, ablation studies, and representative failure cases. The contents are organized as follows:

- **The TimeBlind Benchmark**: data collection procedures, prompts used for data generation, quality control process, benchmark statistics, and taxonomy details.
- **Additional Ablation Studies**: additional ablation experiments on high-FPS settings and across video subsets with different levels of video similarity.
- **Impact Statement**:a discussion of the broader impact of the TimeBlind benchmark.
- **Failure Cases**: representative failure examples (1–7).

A.1 Data Collection Processes

Details on human-recorded videos. We recruited 10 participants and provided each with generated caption pairs and clear instructions: (1) record two videos matching the given captions; (2) maintain static consistency (same people, props, scene, camera angle); (3) ensure temporal minimality (videos differ only in the targeted temporal factor); and (4) meet quality standards (clear footage, similar duration within pairs). Videos are not limited to single-person scenes—some include multi-person interactions across settings such as parks, streets, restaurants, and offices.

Details on internet video collection. Annotators searched YouTube using LLM-generated captions or relevant keywords. They manually identified clips matching caption pairs. Where possible, both clips were sourced from the same source video to maximize static consistency.

Details on generated video collection. We used tools such as Unity to create videos for each caption pair. During generation, we ensured static visual consistency within each video pair, introduced differences only in the targeted temporal aspect, and kept the two videos in each pair nearly the same in duration.

A.2 Quality Control Process

We used 8 annotators for quality review (separate from the 10 video recorders and 4 human evaluators). Each instance was independently reviewed by 2 annotators, and video-question pairs in each instance were accepted only upon unanimous agreement. Instances were kept only if all constituent pairs passed. The inter-annotator agreement was 94.3%, and Cohen's kappa was 0.835. The total review time was 62.2 hours ($\sim$2-3 min per instance per annotator). We also report the per-category annotator agreement scores in Table 8; all three categories achieve substantial agreement ($\kappa > 0.8$). Event Attributes show the lowest agreement, consistent with the inherently more subjective nature of judging continuous properties such as speed and force, but still well above conventional thresholds for high-quality annotation. The final benchmark was then evaluated by 4 separate, independent annotators who achieved 98.2% I-Acc, confirming that it is high-quality and largely unambiguous.

Table 8: Human annotation agreement across categories.

Category	Agreement	Cohen's κ
Events	95.6%	0.88
Event Attributes	93.4%	0.84
Structural Event Logic	93.7%	0.81

A.3 Benchmark Statistics

Data Statistics. TimeBlind includes both two-choice multiple-choice and binary questions in equal numbers. It is a short-video benchmark, with an average video duration of 8.49 seconds. We computed CLIP-based video similarity for all pairs by sampling frames at 1 FPS, extracting per-frame embeddings, and average-pooling within each video. The mean pair similarity is 0.971 (median 0.979), and 88.3% of instances have similarity above 0.95, confirming strong static consistency.

Table 9: **Statistics of the TimeBlind benchmark.**

Statistic	Value
Binary Question	1,200
Two-Choice Multiple-Choice Question	1,200
Avg. Video Length (sec)	8.49
Avg. Question Length (words)	20.54

Statistical significance. We report statistical significance tests below. Wilson 95% confidence intervals on I-Acc ($n = 600$): (1) Random chance: 6.3% [4.6%, 8.5%], (2) Best open-source (Molmo2-8B): 31.2% [27.6%, 35.0%], (3) Best overall (Gemini 3 Pro): 48.2% [44.2%, 52.2%], (4) Human: 98.2% [96.8%, 99.0%].

All intervals are non-overlapping, confirming that performance differences across tiers are statistically significant. Since all models are evaluated on the same 600 instances, we use the paired McNemar's test for pairwise comparisons. The gap between Gemini 3 Pro (48.2%) and human performance (98.2%) is highly significant ($p = 6.69 \times 10^{-81}$). The gap between Gemini 3 Pro and the best open-source model, Molmo2-8B (31.2%), is likewise significant ($p = 1.67 \times 10^{-12}$). A one-sample exact binomial test of Gemini 3 Pro against random chance (6.3%) confirms performance significantly exceeds chance ($p = 1.91 \times 10^{-178}$). Even the small 1.9% gap between GPT-5 (46.3%) and Gemini 3 Pro (48.2%) correctly registers non-significant (McNemar's $p = 0.441$), demonstrating that our sample size has appropriate discriminative power: it captures meaningful gaps while not over-claiming marginal ones.

A.4 Prompt Details.

In this section, we present the full prompt template used in our Human–AI collaborative pipeline (see Section 3.3). The system prompt for TimeBlind consists of three parts: (1) a **General Prompt** that defines the role and task, (2) **Detailed Requirements** that specify the core generation rules and logical constraints, and (3) an **Output Format** that standardizes the response structure. The prompt includes strict structural constraints to ensure high-quality minimal pairs. When using this prompt, we provide a single category as the "Input" each time.

Part 1: Role, Context, & Task Definition

Role. You are an expert in dataset curation and benchmark design for Multimodal Large Language Models (MLLMs).

Context. We are constructing a benchmark called TimeBlind to evaluate the temporal perception capabilities of MLLMs. Each TimeBlind instance consists of two videos (v_1, v_2) and two questions (q_1, q_2).

- The videos are visually similar, sharing the same objects and background, but differ in temporal dynamics.
- The questions are designed to distinguish these specific temporal features.
- During testing, the model receives only **one video** and **one question** at a time.

Task. Given a specific Temporal Dynamic Category, generate a valid data instance consisting of one pair of videos and one pair of questions that satisfies the requirements below.

Part 2: Requirements & Constraints

Requirements.

1. Video Generation (v_1, v_2).

- **Format:** Provide detailed text descriptions for both videos.
- **Feasibility:** The scenes should be easy to collect via web search, self-recording, or synthetic generation.
- **Visual Consistency:** v_1 and v_2 must share the same objects, background, and camera angle. They should look nearly identical in a static frame, with differences only in temporal dynamics.

2. Question Generation (q_1, q_2).

- **Type:** Use either two-choice multiple-choice questions or binary questions (Yes/No).
- **Relevance:** Both questions must relate to the given "Temporal Dynamic Category."
- **Temporal Dependency:** The questions must require temporal understanding and cannot be answered from a single static frame.

Logical Constraints.
The answers A must satisfy the following constraints to ensure that the model is truly perceiving time.

Constraint A: Cross-Video Difference. The answer to each question must change across videos:

$$A(q_1, v_1) \neq A(q_1, v_2),$$
$$A(q_2, v_1) \neq A(q_2, v_2).$$

Constraint B: Intra-Video Difference. The two questions must have different answers within the same video:

$$A(q_1, v_1) \neq A(q_2, v_1),$$
$$A(q_1, v_2) \neq A(q_2, v_2).$$

Part 3: Output Specifications

Output Format.
Present the result using the following structured format.

Temporal Category: [Input Category]

Video Descriptions

- v_1**:** [Description]
- v_2**:** [Description]
- **Collection Source:** [Synthetic / Self-shot / Web]

Questions & Answers

- q_1**:** [Question Text]
  Answer in v_1*:* [Answer]; *Answer in* v_2*:* [Answer]
- q_2**:** [Question Text]
  Answer in v_1*:* [Answer]; *Answer in* v_2*:* [Answer]

Reason

- **Why a single frame is insufficient:** [Brief explanation]

A.5 Taxonomy

In this section, we introduce the detailed taxonomy of TimeBlind as summarized in Table 10.

Table 10: **Hierarchical taxonomy of TimeBlind for video understanding.** The taxonomy is structured into three levels—Event, Event Attribute, and Structural Event Logic. For each level, we provide its conceptual definition, associated subcategories, and representative paired video descriptions.

Taxonomy	Definition	Sub-Categories	Example: Paired Video Descriptions
`Event`	The atomic units of temporal understanding, encompassing fine-grained action phases and diverse state transitions.	**Fine-grained Action** Subtle differences in action phase that are difficult to resolve from a single frame.	• *The man* **opens** *the door.* • *The man* **closes** *the door.*
		State Transitions Tracking attribute changes over time (e.g., color, size, shape).	• *The apple turns* **from red to yellow**. • *The apple turns* **from yellow to red**.
`Event Attribute`	The properties that characterize how an event unfolds over time.	**Speed** The rate at which an action is performed or an object moves over time.	• *The man* **quickly** *opens the door.* • *The man* **slowly** *opens the door.*
		Force The intensity of an action as reflected in visible physical effects (e.g., acceleration, impact strength, or deformation).	• *The man* **gently** *opens the door.* • *The man* **forcefully** *opens the door.*
		Magnitude The spatial extent or scale of motion or deformation.	• *The person waves* **slightly**. • *The person waves* **broadly**.
		Duration The length of time an event lasts.	• *The person holds the cup for a* **long time**. • *The person holds the cup* **briefly**.
		Direction The trajectory or spatial direction of motion.	• *The person moves* **toward the right side of the frame**. • *The person moves* **toward the left side of the frame**.
		Repetition The number of times an event occurs.	• *The person picks up the cup* **three times**. • *The person picks up the cup* **twice**.
`Structural Event Logic`	The higher-level rules governing how events relate to and compose with one another.	**Temporal Topology** Interval relations between event intervals (Allen's interval algebra), such as before, meet, overlap, start, during, finish, and equal.	• *The person picks up the book and cup* **at the same time**. • *The person picks up the book and subsequently picks up the cup* **after a short delay**.
		Causal Contingency The causal dependencies between events.	• *The* **first press** *causes the light to turn off.* • *The* **second press** *causes the light to turn off.*
		Cross-Event Comparison The comparative reasoning over attributes of different events within the same video.	• *The person holds the cup* **longer than** *the book.* • *The person holds the cup for* **less time than** *the book.*

Table 11: **Category-Wise I-Acc (%) on TimeBlind.** This table reports I-Acc for advanced models across 11 fine-grained temporal understanding tasks to pinpoint specific cognitive deficits. Due to space constraints, we use the following abbreviations for temporal categories: ***FG Action***: Fine-Grained Action, ***State Trans***: State Transitions, ***Mag***: Magnitude, ***Dir***: Direction, ***Dur***: Duration, ***Rep***: Repetition, ***Temp Topo***: Temporal Topology, ***Causal Cont***: Causal Contingency, and ***Cross Comp***: Cross-Event Comparison. We include the overall performance (***Avg***) for each high-level category alongside the fine-grained categories. The results show clear performance gaps across categories. Models generally perform well on discrete *Events*, but struggle with *Event Attributes*, which require low-level physical understanding such as *Speed* and *Force*. The best results are **bolded**.

Models	**Event**			**Event Attribute**							**Structural Event Logic**			
	FG Action	State Trans	**Avg**	Speed	Force	Mag	Dir	Dur	Rep	**Avg.**	Temp Topo	Causal Cont	Cross Comp	**Avg**
Random Chance	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3	6.3
Open-Source Models														
InternVL3.5-38B	30.7	33.6	32.4	3.6	0.0	4.1	15.7	50.0	5.6	9.4	10.2	5.0	11.4	9.4
LLaVA-Video-72B	23.9	40.5	33.3	0.0	5.6	0.0	8.6	40.0	11.1	7.3	10.2	10.0	18.2	12.0
Qwen3-VL-235B	38.6	47.4	43.6	3.6	5.6	8.3	18.6	**75.0**	13.9	12.5	**20.4**	7.5	**25.0**	**18.8**
Molmo2-8B	**54.5**	**49.1**	**52.0**	**10.7**	**11.1**	**12.5**	**31.4**	25.0	**19.4**	**20.3**	18.5	**15.0**	18.2	17.7
Closed-Source Models														
GPT-5	**62.5**	**55.2**	**58.3**	21.4	16.7	25.0	**40.0**	**75.0**	**30.6**	32.3	49.1	**50.0**	45.5	48.4
Gemini 3 Pro	49.4	49.0	49.2	**58.3**	**50.0**	**45.0**	23.1	**75.0**	26.5	**36.7**	**69.0**	37.5	**50.0**	**57.8**

A.6 More Analysis in Category-Wise Diagnosis

To pinpoint specific cognitive deficits, we analyze performance across the TimeBlind taxonomy in Table 11.

- **Performance Across Hierarchical Levels.** We observe notable performance gaps across different temporal understanding tasks, with models generally performing better on discrete *Events* (e.g., distinguishing atomic actions) compared to continuous *Event Attributes* or *Structural Event Logic*. For example, GPT-5 achieves 58.3% accuracy in the *Event* category (peaking at 62.5% for *Fine-Grained Action*), but performance declines to 32.3% for *Event Attributes* and 48.4% for *Structural Event Logic*. Furthermore, we note that while LLaVA-Video-72B and InternVL 3.5-38B achieve moderate performance on *Events* (33.3% and 32.4%), they perform only slightly above random chance in the other two categories.
- **Physical Dynamics.** Performance is lowest in the *Event Attributes* category, which requires sensitivity to continuous or qualitative parameters such as kinematics (speed) and dynamics (force, magnitude). Strong proprietary models GPT-5 and Gemini 3 Pro only achieve 32.3% and 36.7% I-Acc, respectively, while the top-performing open-source model, Molmo2-8B, scores only 20.3%. Furthermore, several models, including Qwen3-VL-235B, LLaVA-Video-72B, and InternVL 3.5-38B, perform at near random chance, struggling to distinguish nuances such as *gentle vs. forceful*, *fast vs. slow*, or *large vs. small amplitude*. These results expose a systematic deficiency in current models' understanding of low-level, physics-related temporal dynamics.
- **Gap in Structural Logic.** We observe that GPT-5 and Gemini 3 Pro demonstrate relatively robust performance in *Structural Logic Analysis*, achieving 48.4% and 57.8%, respectively. However, a striking disparity emerges between these proprietary models and their open-source counterparts. The leading Qwen3-VL-235B achieves only 18.8%—trailing Gemini 3 Pro by 39.0%—and scores a mere 7.5% in the *Causal Contingency* sub-category. These results indicate that while open-source MLLMs may effectively detect isolated events, they lack the necessary logical abilities to reason about precise relationships between multiple events, particularly regarding causality.

A.7 More Ablation Study

Number of Input Frames. Table 12 shows that increasing frames from 8 to 32 yields only marginal I-Acc gains: 1–5% for InternVL 3.5 and LLaVA-Video, and less than 2% for GPT-5. Even with sufficient frames, models struggle with TimeBlind, revealing fundamental limitations in fine-grained temporal understanding.

Impact of Model Size. Despite an 10× parameter increase for LLaVA-Video (7B to 72B) and 5× for InternVL 3.5 (8B to 38B), both show less than

10% I-Acc improvement across all frame settings (Table 12). Simply scaling model size does not yield robust spatio-temporal understanding.

Table 12: **Effect of Model Size and Input Frames.** We report the I-Acc performance of several open-source and proprietary models across various model sizes and input frame counts. The results indicate that simply scaling up model size and frame count fails to significantly improve performance. Best results within each model grouping are highlighted in **bold** with a gray background.

Size	Frames	Q-Acc	V-Acc	Acc	I-Acc
InternVL 3.5					
8B	8	34.7	41.5	61.9	17.1
	16	**35.6**	42.7	62.6	**18.5**
	32	**35.6**	**43.1**	**62.9**	17.4
14B	8	34.2	46.3	63.6	19.9
	16	**38.2**	47.6	65.2	23.5
	32	37.7	**49.4**	**65.4**	**23.9**
38B	8	35.8	42.8	63.2	20.4
	16	38.9	45.4	64.5	24.4
	32	**39.4**	**46.4**	**65.9**	**25.1**
LLaVA-Video					
7B	8	21.8	37.9	59.5	12.4
	16	23.4	39.8	60.1	13.9
	32	**24.7**	**41.1**	**60.7**	**14.6**
72B	8	29.7	46.2	63.5	19.9
	16	32.4	48.4	65.0	23.0
	32	**32.8**	**48.7**	**65.2**	**23.3**
GPT-5					
unknown	8	60.9	69.2	79.0	49.1
	16	**62.7**	**69.9**	**79.5**	**50.9**
	32	61.1	68.2	78.7	48.3

Video Similarity. We computed CLIP-based video similarity for all video pairs by sampling frames at 1 FPS, extracting frame-level embeddings, and average-pooling them within each video. We further evaluate GPT-5 on two similarity-based subsets, namely the top 30% highest-similarity pairs and the bottom 30% lowest-similarity pairs. As shown in Table 13, GPT-5's performance is comparable across the highest- and lowest-similarity subsets (45.7% vs 47.8%), indicating that the benchmark's difficulty does not depend strongly on residual visual differences between paired videos. Combined with the shortcut analysis in Table 4 (single-frame, shuffled-frame, and language-only baselines all near random chance), this confirms that TimeBlind is robust to subtle residual visual cues.

A.8 Impact Statement

This work introduces a diagnostic benchmark for evaluating temporal understanding in video-language models. We anticipate several positive impacts: (1) TimeBlind can guide development of more temporally-aware models, which is critical for applications in robotics, autonomous driving, and assistive technologies; (2) by revealing systematic failure modes, our benchmark may help practitioners avoid deploying models in safety-critical scenarios where temporal reasoning is essential.

We also acknowledge potential concerns. Improved video understanding capabilities could enable more sophisticated surveillance or content moderation systems, raising privacy considerations. Additionally, our benchmark primarily features videos from controlled settings and internet sources, which may not represent the full diversity of real-world scenarios or populations. We encourage future work to expand temporal reasoning evaluation to more diverse contexts.

The benchmark itself does not enable direct harmful applications, as it evaluates rather than enhances model capabilities. We will release our data and code to support reproducible research.

Table 13: **GPT-5 Performance on different similarity-based subsets.** The table reports GPT-5's I-Acc results on the high-similarity subset, the low-similarity subset, and the full benchmark. GPT-5's performance is comparable across the highest- and lowest-similarity subsets (45.7% vs 47.8%), indicating that the benchmark's difficulty does not depend strongly on residual visual differences between paired videos.

Subset	I-Acc
Top 30% (highest similarity)	45.7
Bottom 30% (lowest similarity)	47.8
Full benchmark	46.3

A.9 License and Intended Use

License. We release the TimeBlind benchmark—including all curated video–question pairs, annotations, and evaluation code—under the Creative Commons Attribution 4.0 International License (CC BY 4.0)[1]. For videos sourced from the internet, we retain the original license terms of each source. For videos recorded by participants and generated via simulation, we release them under CC BY 4.0.

Intended Use. TimeBlind is intended solely as a diagnostic research benchmark for evaluating the compositional spatio-temporal understanding capabilities of multimodal large language models. It is

[1] https://creativecommons.org/licenses/by/4.0/

not designed for training, fine-tuning, or deployment in production systems, and should not be used as a basis for safety-critical decisions involving real-world video analysis.

Use of Existing Artifacts. Our use of existing artifacts is consistent with their intended use. We evaluate publicly released MLLMs (e.g., GPT-5, Gemini 3 Pro, Qwen3-VL) in zero-shot inference settings consistent with their official APIs and documented evaluation protocols. Internet-sourced videos were collected from publicly accessible platforms (e.g., YouTube) for non-commercial research purposes, in line with platform terms of service. Synthetic videos generated via simulation tools (e.g., Unity) are produced and distributed in accordance with the tools' research-use terms.

A.10 Failure Cases

In this section, we present failure cases for state-of-the-art models, including closed-source models (GPT-5 and Gemini 3 Pro) and open-source models (Qwen3-VL-235B and Molmo2-8B). For each failure case, we show one question, two videos with their corresponding captions, and each model's answer to the question for both videos. We indicate whether the answer is correct or incorrect using check marks and cross marks.

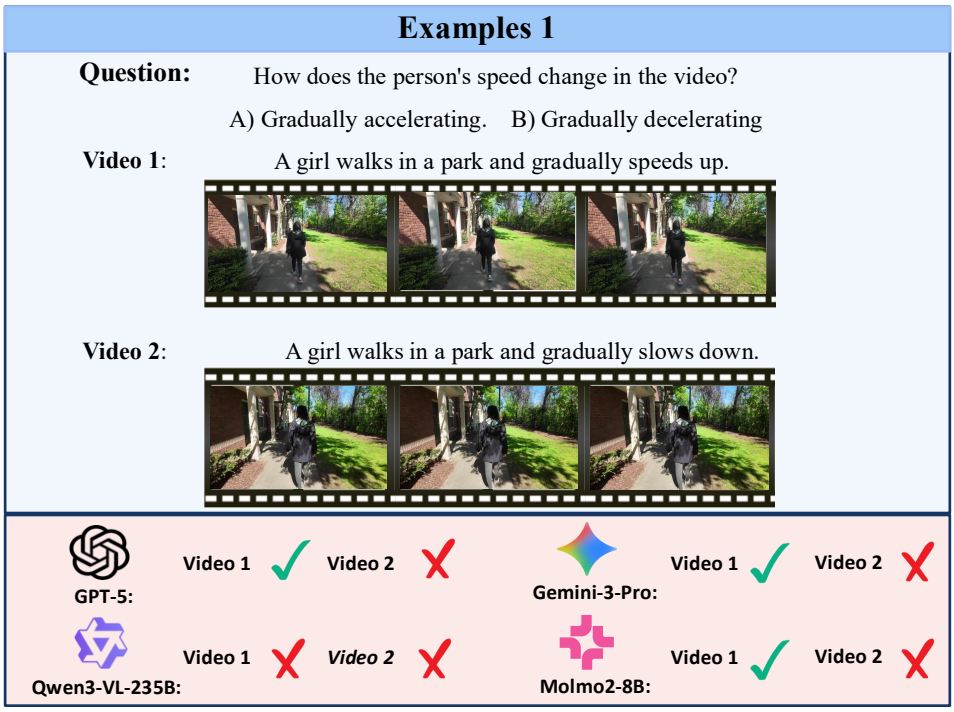

Figure 4: Failure Case 1

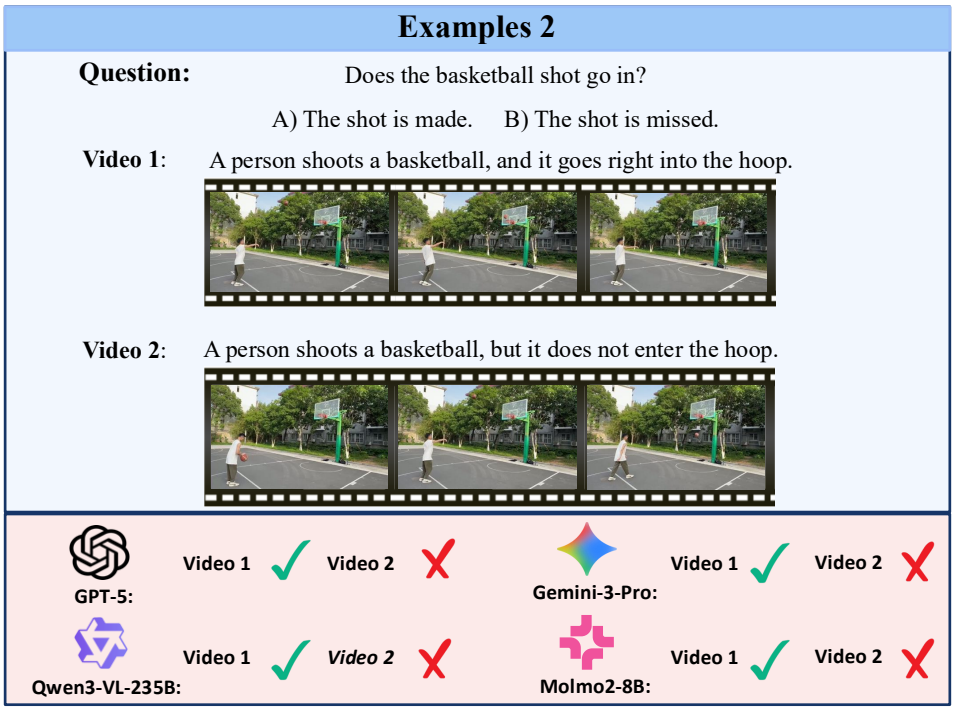

Figure 5: Failure Case 2

Examples 3

Question:

Do the bottom lights turn on after a pause once the top lights go out?

Video 1: There is a small time gap between the top lights turning off and the bottom lights turning on.

Video 2: The bottom lights turn on as soon as the top lights turn off.

GPT-5:	Video 1 ✓	Video 2 ✓		Gemini-3-Pro:	Video 1 ✓	Video 2 ✓	
Qwen3-VL-235B:	Video 1 ✗	Video 2 ✗		Molmo2-8B:	Video 1 ✓	Video 2 ✗	

Figure 6: Failure Case 3

Examples 4

Question:

Does the number of blue darts on the dartboard decrease?

Video 1: The number of blue darts on the dartboard increases.

Video 2: The number of blue darts on the dartboard decreases.

GPT-5:	Video 1 ✓	Video 2 ✓		Gemini-3-Pro:	Video 1 ✓	Video 2 ✓	
Qwen3-VL-235B:	Video 1 ✓	Video 2 ✗		Molmo2-8B:	Video 1 ✗	Video 2 ✓	

Figure 7: Failure Case 4

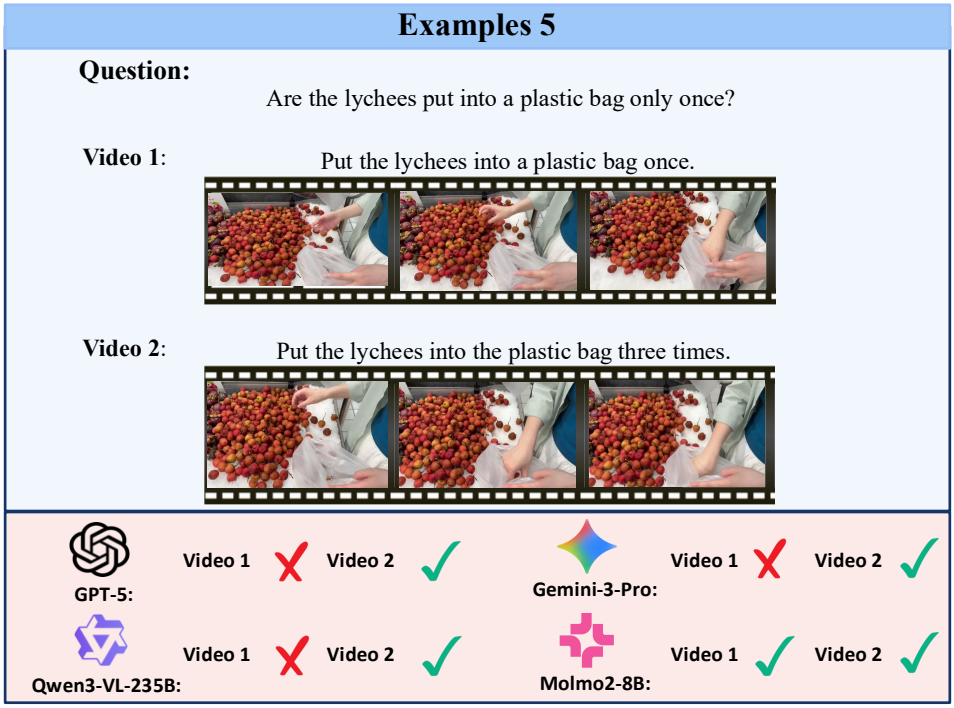

Figure 8: Failure Case 5

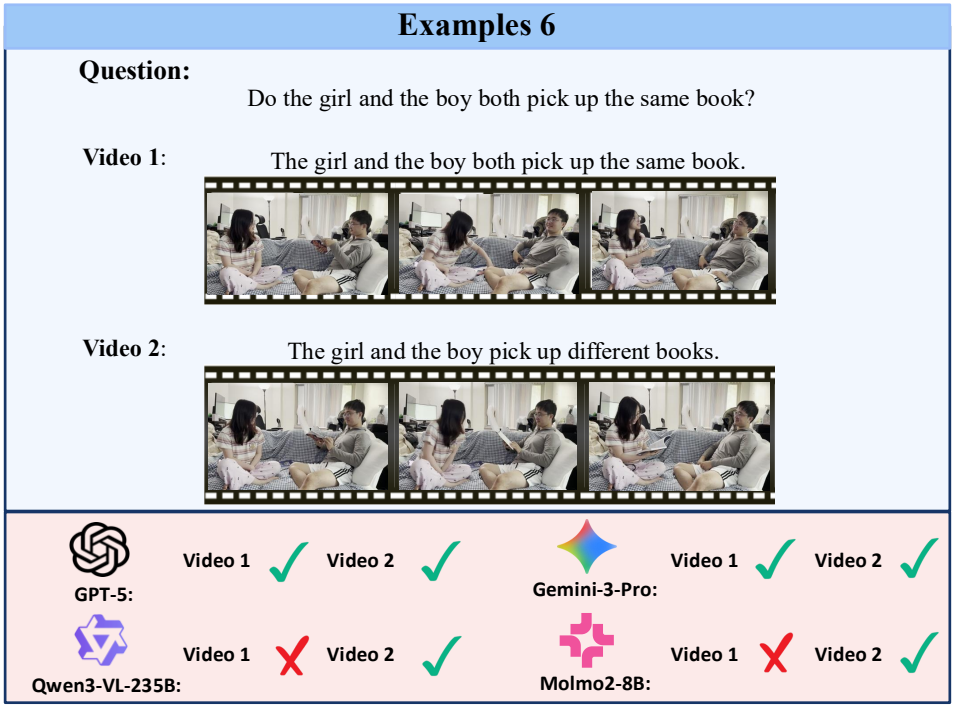

Figure 9: Failure Case 6

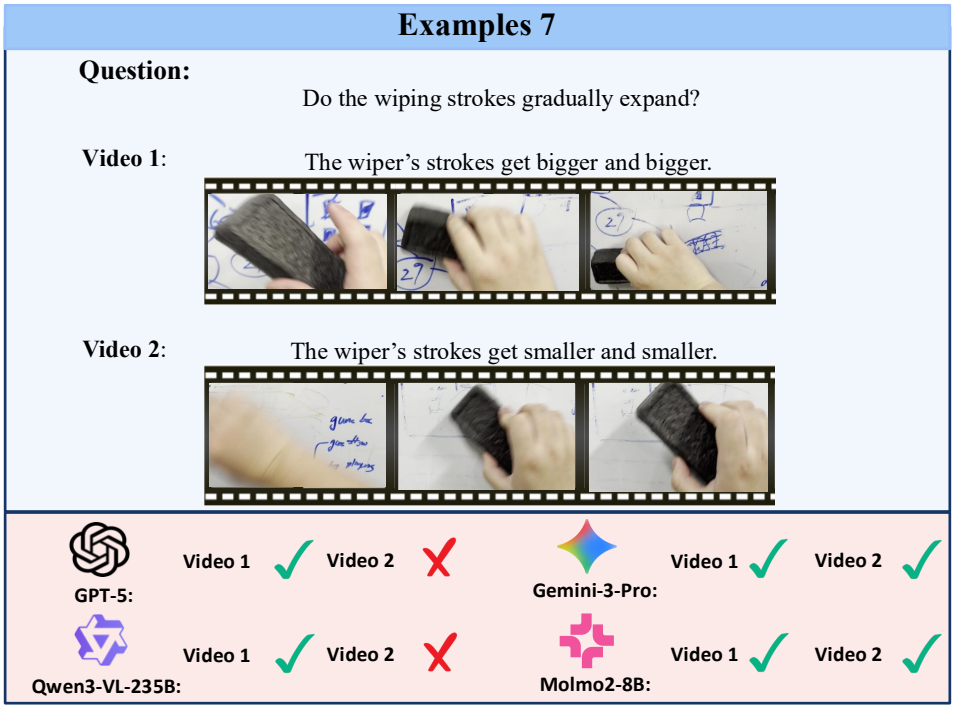

Figure 10: Failure Case 7