CausalStep: A Benchmark for Explicit Stepwise Causal Reasoning in Videos

Xuchen Li[1,2,3*], **Xuzhao Li**[4*], **Shiyu Hu**[4], **Kaiqi Huang**[1,2†], **Wentao Zhang**[3,5†]
[1]Institute of Automation, Chinese Academy of Sciences
[2]University of Chinese Academy of Sciences
[3]Zhongguancun Academy
[4]Nanyang Technological University
[5]Peking University
s-lxc24@bjzgca.edu.cn, xuzhaoli2001@gmail.com, kqhuang@ia.ac.cn, wentao.zhang@pku.edu.cn

Abstract

Recent advances in large language models (LLMs) have improved reasoning in text and image domains, yet achieving robust video reasoning remains a significant challenge. Existing video benchmarks mainly assess shallow understanding and reasoning and allow models to exploit global context, failing to rigorously evaluate true causal and stepwise reasoning. We present CausalStep, a benchmark designed for explicit stepwise causal reasoning in videos. CausalStep segments videos into causally linked units and enforces a strict stepwise question-answer (QA) protocol, requiring sequential answers and preventing shortcut solutions. Each question includes carefully constructed distractors based on error type taxonomy to ensure diagnostic value. The benchmark features 100 videos across six categories and 1,852 multiple-choice QA pairs. We introduce seven diagnostic metrics for comprehensive evaluation, enabling precise diagnosis of causal reasoning capabilities. Experiments with leading proprietary and open-source models, as well as human baselines, reveal a significant gap between current models and human-level stepwise reasoning. CausalStep provides a rigorous benchmark to drive progress in robust and interpretable video reasoning.

Introduction

Recent advances in large language models (LLMs) have driven impressive progress in text (Jiang and Li 2024), image (Zhang et al. 2023), and general video understanding (Tang et al. 2025). However, extending these reasoning capabilities to complex, real-world video scenarios (Wang et al. 2025) remains a major challenge. Video reasoning is fundamentally different from text or static images, as videos encode rich, sequential, and multimodal information that requires models to perform long-range, multi-frame reasoning and evidence integration across both temporal and spatial dimensions. This capability is essential for applications such as embodied intelligence (Roy et al. 2021), intelligent surveillance (Ibrahim 2016), and human-computer interaction (MacKenzie 2024).

Despite recent progress, existing video reasoning benchmarks (Li et al. 2024b; Liu et al. 2024a; Hu et al. 2025; Zhao et al. 2025; Zhu et al. 2025b; Cheng et al. 2025) exhibit key limitations. Most benchmarks focus on perception or shallow understanding, requiring only the identification of relevant frames or context. Crucially, by typically providing the entire video as input, these benchmarks allow models to exploit global information or shortcut strategies, thereby failing to assess true causal and stepwise reasoning. As a result, they do not capture the causally grounded reasoning processes humans naturally employ when interpreting complex video narratives. Moreover, the design of distractor options in multiple-choice questions is often unsystematic, lacking systematic coverage of common reasoning errors and thus failing to rigorously challenge model robustness.

To address these gaps, we introduce **CausalStep**, a new benchmark specifically designed to evaluate *explicit stepwise causal reasoning* in videos. In CausalStep, each video is manually segmented into a sequence of causally linked segments. At each step, the model is provided with the current segment along with the previous segment (if any), without access to future information, and must answer a question—either a descriptive understanding question or an explicit causal reasoning question—before it can access the next. This protocol strictly enforces sequential, causally dependent reasoning and precludes the use of global shortcuts. Furthermore, we design a novel distractor generation strategy: for each multiple-choice question, distractor options are systematically constructed according to a taxonomy of error types, including temporal confusion, causal misattribution, and object misrecognition. This ensures each question not only evaluates surface-level distractor but also challenges the model's ability to distinguish between plausible but incorrect alternatives.

CausalStep comprises 100 videos spanning six diverse categories (e.g., cartoons, movies, sports, performances, documentaries, and TV shows), totaling 1,852 multiple-choice question-answer (QA) pairs. Each question is carefully annotated and reviewed, covering both descriptive understanding and explicit stepwise causal reasoning tasks—enabling fine-grained analysis of models' causal reasoning abilities. To provide a comprehensive assessment of model performance, we propose a suite of seven diagnos-

[*]These authors contributed equally.
[†]Corresponding Authors.
Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Benchmark	#Videos	Duration	#QA Pairs	Spatio-temporal	Causal	Stepwise	Annotation
MVBench	200	15-20 s	4,000	✓	✓	X	A
TempCompass	410	15-20 s	7,540	✓	X	X	A&M
Video-MMMU	300	506.2 s	900	X	X	X	M
MMVU	1,529	51.4 s	3,000	X	X	X	M
Video-MME	900	35.7 s	1,944	✓	✓	X	M
VCR-Bench	859	159 s	1,034	✓	✓	X	A&M
Video-Holmes	270	160 s	1,837	✓	✓	X	A&M
MMR-V	317	277 s	1,257	✓	✓	X	A&M
Ours	**100**	**430.5 s**	**1,852**	✓	✓	✓	A&M

Table 1: Comparison between CausalStep and existing video understanding/reasoning benchmarks across key aspects: the number of videos (**#Videos**), video duration (**Duration**), number of reasoning QA pairs (**#QA Pairs**), spatio-temporal relationship understanding (**Spatio-temporal**), causal relationship understanding (**Causal**), stepwise reasoning protocol (**Stepwise**), and annotation methodology (**Annotation**). **A** denotes AI-generated, **M** denotes manual, and **A&M** indicates a combination.

tic metrics: chain success rate, average and maximum chain length, restart frequency, weighted score, and dedicated accuracies for descriptive understanding and isolated causal reasoning. These metrics capture not only overall accuracy but also the depth, stability, and robustness of a model's reasoning process.

We conduct extensive experiments on CausalStep, evaluating a wide range of state-of-the-art proprietary and open-source multimodal models—including the latest GPT (OpenAI 2025, 2024b,a), Gemini (Reid et al. 2024; Google DeepMind 2025), Claude (Anthropic 2024), Qwen (Yang, Yang et al. 2024), Gemma (Kamath, Ferret et al. 2025), InternVL (Chen et al. 2024; Zhu et al. 2025a), LLaVA (Li et al. 2024a; Zhang et al. 2024; Lin et al. 2023), and Phi (Abouelenin, Ashfaq et al. 2025) series—as well as human participants. Our results reveal a substantial gap between current models and human-level performance, especially in explicit stepwise causal reasoning. This disparity is primarily driven by models' difficulty in maintaining continuous, error-free reasoning chains and their vulnerability to subtle distractors. These results show that even the strongest models struggle with long-range causal integration and are susceptible to confusable distractors, highlighting the need for further advances in video reasoning of multimodal large language models (MLLMs).

Our main contributions are as follows:

- **A novel benchmark for explicit stepwise causal reasoning in videos:** We introduce CausalStep, which segments videos into causally linked units and enforces a strict stepwise QA protocol, enabling rigorous evaluation of sequential, causally grounded reasoning in complex video narratives.
- **A comprehensive annotation and evaluation framework:** We design a hybrid annotation pipeline combining LLM generation and human review, and propose a taxonomy-based distractor generation strategy. We further introduce seven diagnostic metrics that provide a fine-grained, multi-dimensional assessment of model performance, covering reasoning depth and robustness.
- **Extensive empirical analysis and insights:** We benchmark a diverse set of state-of-the-art (SOTA) proprietary and open-source models, as well as human baseline on CausalStep. Our experiments reveal a significant gap between current models and human-level stepwise causal reasoning, and provide actionable insights for future research on robust and interpretable video reasoning.

Related Work

MLLMs for Video Understanding and Reasoning

Recent progress in MLLMs has significantly advanced the field of video understanding and reasoning (Tang et al. 2025; Wang et al. 2025; Cao et al. 2025; Li et al. 2025). Building on breakthroughs in image-based multimodal reasoning, models such as Gemma (Kamath, Ferret et al. 2025), LLaVA-Onevision (Li et al. 2024a), Phi (Abouelenin, Ashfaq et al. 2025), InternVL (Chen et al. 2024), Video-LLaVA (Zhang et al. 2024), Qwen-VL (Yang, Yang et al. 2024), GPT series (OpenAI 2024b), Claude (Anthropic 2024) and Gemini (Reid et al. 2024) have adapted LLMs to process videos as sequences of frames, enabling temporal and contextual analysis. This typically involves utilizing visual encoders to extract frame features, which are then integrated with linguistic components. Despite these advancements, current MLLMs still face challenges in performing long-range, stepwise causal reasoning, especially when required to integrate information across multiple frames.

Video Understanding Benchmark

Video understanding benchmarks (Tang et al. 2025; Hu et al. 2024; Li et al. 2024c,d,e) have evolved from early datasets on basic perception—like action recognition (Kong and Fu 2022; Li et al. 2024f) and short-clip QA (e.g., MSRVTT-QA (Xu et al. 2017))—to broader evaluations. Recent benchmarks cover diverse content, tasks, and longer videos. Video-MME (Fu et al. 2025) and MVBench (Li et al. 2024b) expand to multiple task formats and longer videos. LVBench (Wang et al. 2024b) and LongVideoBench (Wu et al. 2024b) introduce long-form video QA, while MLVU (Zhou et al. 2024) offers multi-task long-video understanding. E.T. Bench (Liu et al. 2024b) targets open-ended

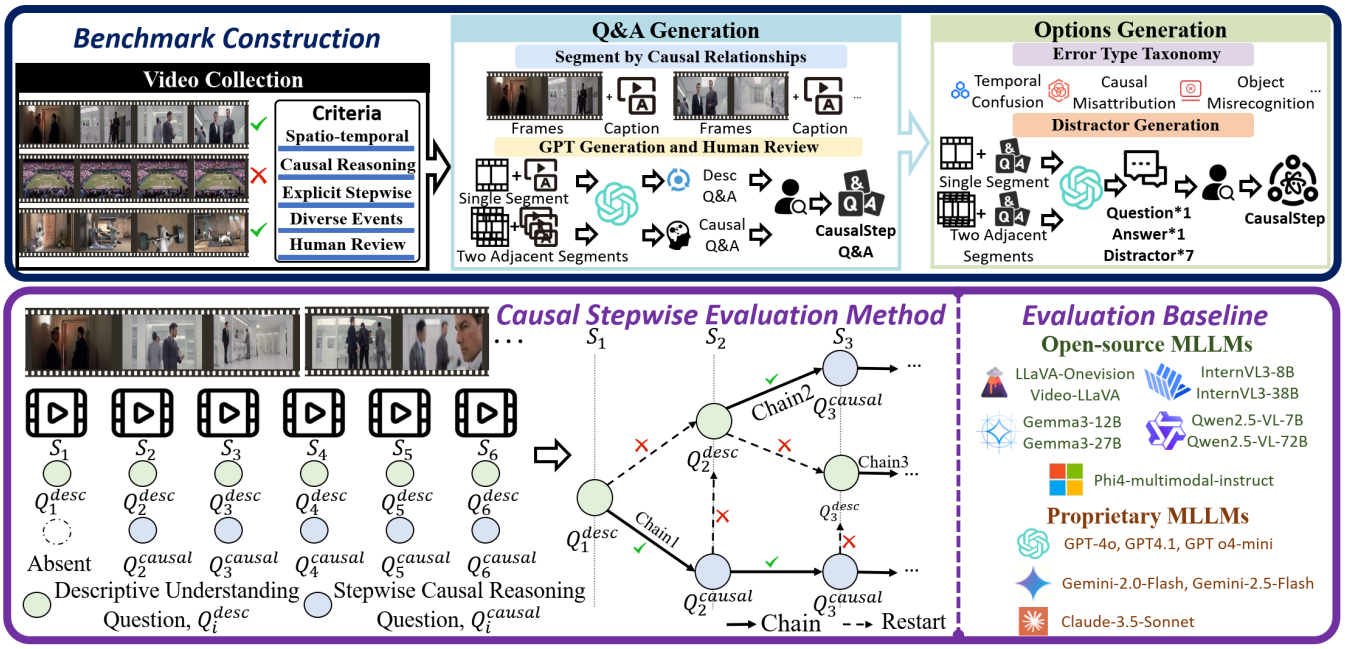

Figure 1: Overview of the CausalStep Benchmark Construction and Evaluation Framework. The top panel illustrates the benchmark construction pipeline: **Video Collection** with specific filtering, **Q&A Generation** leveraging GPT-4o and human review for descriptive and causal questions, and **Options Generation** using a novel error type taxonomy. The bottom panel details the **Evaluation Method**, employing a strictly sequential, stepwise QA protocol. It also visualizes the **Causal Stepwise Evaluation** process with its enforced chain dependencies and restart mechanism, and lists the **Evaluation Baseline** models (open-source and proprietary MLLMs) used in our experiments.

event-level reasoning, and TemporalBench (Cai et al. 2024) evaluates fine-grained temporal analysis. InstructionBench (Wei et al. 2025) tests step-by-step instructional videos for event/object-level reasoning. However, these benchmarks primarily evaluate perceptual and intuitive understanding, often not requiring complex, multi-step reasoning or integration of information across distant frames.

Video Reasoning Benchmark

The push toward more challenging video reasoning tasks (Fei et al. 2024) has led to the development of specialized benchmarks that go beyond perception. TempCompass (Liu et al. 2024a) and MVBench (Li et al. 2024b) focus on temporal and implicit causal reasoning in videos. In parallel, MMVU (Zhao et al. 2025) and VideoMMMU (Hu et al. 2025) focus on reasoning in scientific or educational contexts, while VSI-Bench (Yang et al. 2025) targets indoor scene reasoning. Earlier work like VCR-Bench (Qi et al. 2025) pioneered explicit chain-of-thought reasoning in videos, whereas STAR (Wu et al. 2024a) and VideoVista (Li et al. 2024g) evaluate situated reasoning and versatile video QA, respectively. Adding depth, SOK-Bench (Wang et al. 2024a) integrates open-world knowledge. Video-Holmes (Cheng et al. 2025) further decomposes spatio-temporal reasoning and multi-clue integration across a video. Nevertheless, there remains a gap in benchmarks that rigorously evaluate explicit, stepwise causal reasoning across long video sequences—a gap that CausalStep aims to address.

CausalStep Task Overview

The overview of CausalStep is illustrated in Figure 1. In CausalStep, we propose an explicit stepwise causal reasoning task designed to rigorously evaluate a model's ability to perform human-like, sequential causal reasoning over video content. The task is defined by the following components:

Causal Segmentation of Videos. Each video is manually segmented into a sequence of causal segments, based on its underlying narrative and event structure in MGIT (Hu et al. 2023). A video is thus divided into N segments, denoted as $\{S_1, S_2, \ldots, S_N\}$, where each segment corresponds to a distinct causal event or state. This segmentation is designed to support subsequent stepwise causal reasoning.

Question Types. For every segment, we annotate a descriptive understanding question (Q_i^{desc}) that assesses the model's comprehension of the observable content. These descriptive questions serve as the foundational starting point for each reasoning chain. For each segment except the first (S_i, $i > 1$), we further annotate an explicit stepwise causal reasoning question (Q_i^{causal}), which requires the model to reason about the causal relationship between the current segment (S_i) and its direct preceding segment (S_{i-1}). Note that the first segment (S_1) does not have a causal reasoning question.

Stepwise Reasoning Chain. The reasoning chain for evaluation begins with the descriptive understanding QA for

Algorithm 1: CausalStep Evaluation Framework

Input:
Segments $[S_1, S_2, \ldots, S_N]$;
Descriptive QA list $[Q_1^{desc}, Q_2^{desc}, \ldots, Q_N^{desc}]$;
Reasoning QA list $[Q_2^{causal}, \ldots, Q_N^{causal}]$;
Model `M`
Output: Total score for the video

```
score ← 0;
chain_length ← 0;
i ← 1;
current_question_type ← 'desc';
while i ≤ N do
    if current_question_type == 'desc' then
        desc_ans ← M.Ans(Q_i^desc, S_i)
        if is_correct(desc_ans) then
            chain_length ← chain_length + 1;
            score ← score + 1;
            i ← i + 1;
            current_question_type ← 'causal';
        end
        else
            chain_length ← 0;              // Restart
            i ← i + 1;
            current_question_type ← 'desc';
        end
    end
    if current_question_type == 'causal' then
        if i > N then
            break;
        end
        causal_ans ← M.Ans(Q_i^causal, A_{i-1}, [S_{i-1}, S_i])
        if is_correct(causal_ans) then
            chain_length ← chain_length + 1;
            score ← score + chain_length;
            i ← i + 1;
            current_question_type ← 'causal';
        end
        else
            chain_length ← 0;              // Restart
            i ← i + 1;
            current_question_type ← 'desc';
        end
    end
end
return: score;
```

the first segment (S_1). If the current descriptive understanding question (Q_i^{desc}) is answered correctly, the model will receive a score for that question, and the evaluation chain will immediately proceed to the explicit stepwise causal reasoning QA in the subsequent segment (S_{i+1}). If any answer is incorrect, the reasoning chain is interrupted (see Restart Mechanism). At each step with a causal reasoning QA (Q_i^{causal}), the model is provided with the current segment (S_i) and its direct preceding segment (S_{i-1}), along with its previously correct answer. It cannot access future segments or questions in advance, strictly enforcing a stepwise progression and dependence on prior correct inferences.

Restart Mechanism. If the model answers the current explicit stepwise causal reasoning QA incorrectly (e.g., at segment S_k), the current reasoning chain is interrupted, and it must restart from the descriptive understanding QA of the same segment (S_k). Whereas if the model answers any descriptive understanding QA incorrectly, the reasoning chain is interrupted, and it must restart from the descriptive understanding QA of the next segment (S_{k+1}) to initiate a new reasoning chain. This mechanism prevents accidental progression and ensures validity at each step of a successful causal chain.

Scoring Scheme. Each correct descriptive understanding question is assigned a fixed score of 1 point. Each correct causal reasoning question's score is tied to its position within the current uninterrupted reasoning chain: the first causal question in a chain is worth 1 point, the second 2 points, and so on. If a reasoning chain is interrupted and restarted, the scoring for subsequent causal questions resets to 1 for the newly initiated chain. This mechanism can evaluate the model's ability to maintain longer correct reasoning sequences and provides a fine-grained measure of its explicit stepwise causal reasoning ability.

This task design establishes a rigorous protocol for evaluating models on explicit, sequential, and causally grounded reasoning, preventing shortcut solutions and mirroring human stepwise understanding. Algorithm 1 details the full evaluation process, ensuring transparency and reproducibility.

CausalStep Benchmark Construction

Video Collection

Video Data Sourcing and Filtering. CausalStep draws inspiration from the recently proposed MGIT (Hu et al. 2023) benchmark for its video data collection. MGIT references film narrative principles for video selection, focusing on causal event changes across temporal and spatial dimensions. It comprises 150 long video sequences with rich spatio-temporal and causal relationships, manually annotated at three semantic granularities (action, activity, and story). While MGIT itself is not designed for stepwise reasoning, its action-level segmentation specifically provides a foundation for constructing our explicit stepwise causal reasoning tasks.

Building upon MGIT, we curate a subset of videos specifically for stepwise causal reasoning evaluation. Our filtering prioritizes videos that: (1) support explicit stepwise causal reasoning with interconnected events across temporal segments; (2) discourage shortcut solutions where answers can be inferred from a single scene; and (3) feature key events distributed across different times and/or locations to prevent reliance on local context. This results in a more challenging foundation for CausalStep. We retain 100 videos, averaging 430.5 seconds each, spanning six diverse categories: Cartoons, Movies & TV Shows, Outdoor Sports, Regular Sports, Performances, and Documentaries, ensuring broad coverage of real-world scenarios for explicit stepwise causal

reasoning.

Video Annotation

We employ a hybrid annotation process, combining the efficiency of LLMs with the quality control of human review, to generate high-fidelity QA pairs. For the relevant prompts and human review principles, please refer to Appendix B and C.

Question and Answer Generation. For each video, we segment it based on MGIT's action-level annotations, ensuring boundaries align with genuine causal transitions. This guarantees each segment reflects a distinct causal event or state change, foundational for stepwise reasoning. Detailed segment descriptions are input to GPT-4o (OpenAI 2024b), which generates candidate QA pairs for both descriptive understanding and explicit stepwise causal reasoning. Prompts are designed for diversity, clarity, and task alignment. Human annotators then meticulously review and refine all candidate QA pairs, filtering ambiguous or low-quality items, and ensure factual accuracy and proper grounding in the video segments and causal chains. This two-stage process leverages LLM efficiency while maintaining high annotation quality and task validity.

Taxonomy-Based Distractor Generation. For each multiple-choice question, the correct answer comes from our refined QA pair. To ensure consistent difficulty and introduce explicit "error type" design, we propose a novel taxonomy-based distractor generation approach. We first define typical error types (e.g., temporal confusion, causal misattribution, object misrecognition). Distractor options are then systematically generated by GPT-4o (OpenAI 2024b) to be plausible, incorrect, contextually relevant, and semantically similar alternatives, specifically aligned with these error types. Human annotators meticulously review and edit these distractors, ensuring they are non-trivial, factually sound, and maintain comparable plausibility while fitting their intended error type. Option order is randomized during evaluation to mitigate positional bias. This process maximizes challenge, prevents models from relying on superficial cues, and makes distractors diverse and diagnostically informative, enhancing the benchmark's rigor. Figure 2 shows a QA pair example; Appendix A details the error types.

Benchmark Statistics

CausalStep is a comprehensive benchmark comprising 100 videos (average duration 430.5 seconds, ranging from 149 to 994.4 seconds) across 6 diverse categories. Each video is meticulously segmented into an average of 8.76 causal segments (ranging from 2 to 51 segments per video), forming the basis for our stepwise reasoning tasks. The benchmark features a total of 1,852 multiple-choice QA pairs, evenly split between 926 descriptive understanding questions and 926 causal reasoning questions. Each question averages 8 options, including 1 correct answer and 7 challenging distractors meticulously designed using a novel error-type taxonomy. The entire annotation process employs a hybrid AI-assisted and manual review approach to ensure high data

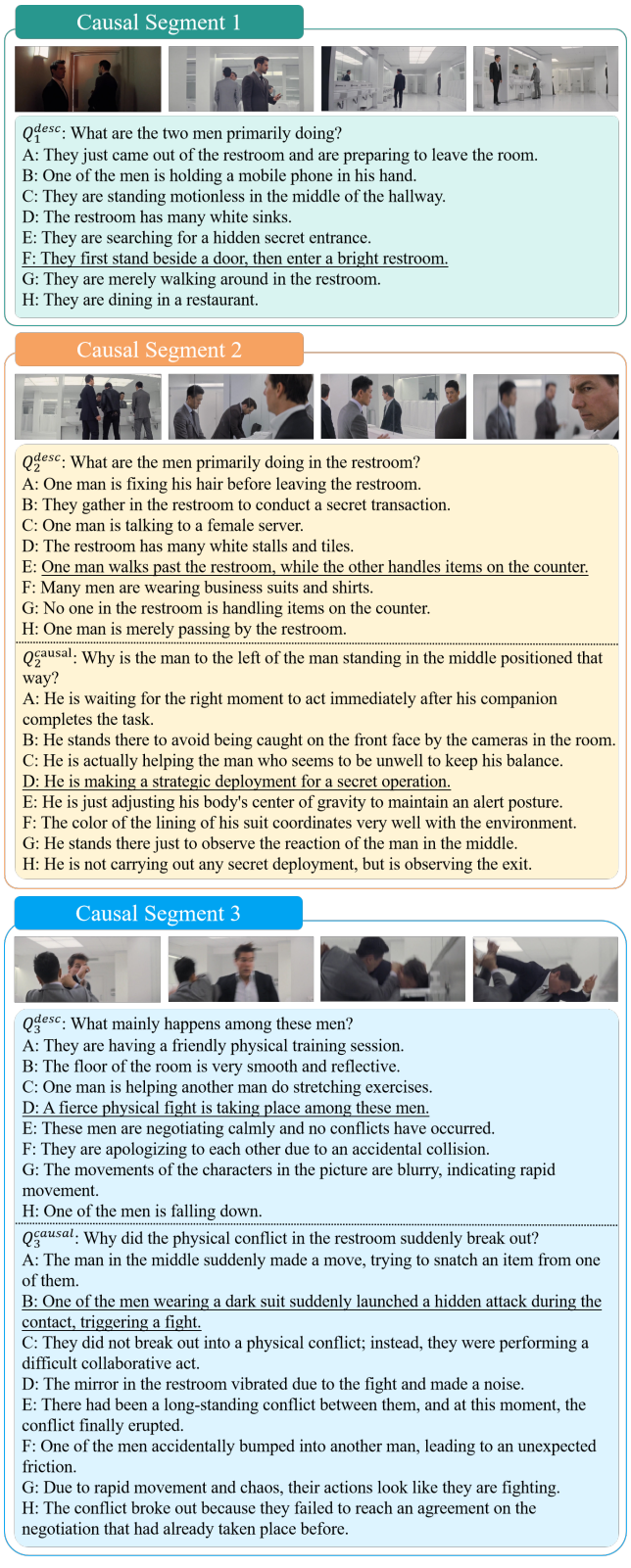

Figure 2: Example of the QA pairs in the CausalStep benchmark. The figure shows three consecutive causal segments, along with their corresponding Descriptive Understanding Q_i^{desc} and Stepwise Causal Reasoning Q_i^{causal}.

Statistic	Value
#Videos	100
Video duration (mean)	430.5 s
Video duration (min / max)	149 s / 994.4 s
#QA pairs	1,852
QA type	Multiple-choice
Options per question	8
#Categories	6
Avg. segments per video	8.76
Segments per video (min / max)	2 / 51
Annotation	AI-assisted + Manual
Distractor design	Error-type taxonomy
Descriptive QA pairs	926
Reasoning QA pairs	926

Table 2: Statistics of the CausalStep benchmark.

quality. Key statistics are summarized in Table 2, providing a detailed overview of the benchmark's scale and characteristics. For more statistical information about the CausalStep benchmark, please refer to Appendix D.

Experiments

Settings

Baselines. We evaluate a comprehensive set of models, including both leading proprietary and open-source MLLMs. Specifically, the closed-source baselines include the GPT series (GPT-4o (OpenAI 2024b), GPT-4.1-2025-04-14 (OpenAI 2025) and o4-mini-2025-04-16 (OpenAI 2025)), the Gemini series (Gemini-2.0-Flash, Gemini-2.0-Flash-thinking (Reid et al. 2024), Gemini-2.5-Flash (Google DeepMind 2025)), Claude-3.5-Sonnet-20241022 (Anthropic 2024). The open-source baselines include the Qwen series (Qwen2.5-VL 7B/72B-Instruct (Yang, Yang et al. 2024)), the Gemma series (Gemma3 12B/27B (Kamath, Ferret et al. 2025)), the InternVL series (InternVL3 8B/38B (Zhu et al. 2025a)), the LLaVA series (LLaVA-OneVision-7B (Li et al. 2024a), Video-LLaVA-7B (Lin et al. 2023)), and Phi4-multimodal-Instruct (Abouelenin, Ashfaq et al. 2025).

For a fair comparison, all models use consistent video frame sampling strategies and input formats, including the same number of frames. Implementation details and prompts are in Appendix E. We also conduct human experiments to establish an upper bound and analyze the human-MLLM reasoning gap. More details please refer to Appendix F.

Metrics. To comprehensively evaluate model performance on explicit stepwise causal reasoning, we employ five key metrics and two supplementary indicators:

- **Chain Success Rate (CSR)**: Proportion of videos where the model completes the entire reasoning chain without errors, reflecting global consistency and long-range reasoning ability.
- **Average Maximum Chain Length (AMCL)**: Average length of the longest uninterrupted reasoning chains achieved across all videos, indicating typical reasoning depth.
- **Maximum Chain Length (MCL)**: The single longest uninterrupted reasoning chain achieved in any video, denoting peak reasoning depth.
- **Restart Frequency (RF)**: How often reasoning chains are interrupted and restarted due to incorrect answers; lower RF indicates greater robustness.
- **Weighted Score (WS)**: Scores increase for later steps in a correct chain (as detailed in task, rewarding longer, sustained reasoning sequences.

Two supplementary indicators further dissect model capabilities:

- **Descriptive Understanding Accuracy (DUA)**: Accuracy on isolated descriptive questions for each segment, measuring foundational visual perception.
- **Isolated Causal Reasoning Accuracy (ICRA)**: Accuracy on causal questions when only the current segment is provided, revealing reliance on local evidence for causal reasoning.

Main Results

Table 3 summarizes the performance of a diverse set of open-source and proprietary models, alongside human baselines, on CausalStep using seven diagnostic metrics. Overall, we observe a clear stratification of model capabilities. Proprietary models consistently outperform open-source models across all metrics, with o4-mini-2025-04-16 achieving the best performance among all evaluated models. Specifically, o4-mini attains a Chain Success Rate (CSR) of 51%, an Average Maximum Chain Length (AMCL) of 7.19, a Maximum Chain Length (MCL) of 30, and a Weighted Score (WS) of 55.06, while maintaining a relatively low Restart Frequency (RF) of 1.69. In contrast, the top open-source model, Gemma-3-27b-it, lags significantly with a CSR of 29%, AMCL of 5.94, and MCL of 20.

Human participants set a strong performance ceiling, achieving a CSR of 79%, AMCL of 8.03, MCL of 46, the highest WS, and the lowest RF among all evaluated entities. For the single-step metrics, Descriptive Understanding Accuracy (DUA) and Isolated Causal Reasoning Accuracy (ICRA), proprietary models again show superior performance over open-source counterparts. However, all models demonstrate a substantial performance gap compared to human participants, particularly evident in ICRA, highlighting the inherent challenge of causal reasoning without the benefit of full contextual understanding derived from a successful chain.

Analysis

MLLMs' Strengths and Limitations in CausalStep. Our findings reveal a persistent gap between current MLLMs and human-level performance across all diagnostic metrics, underscoring the demanding nature of the CausalStep benchmark. While the best-performing model, o4-mini, achieves a CSR of 51% and a MCL of 30, human participants reach significantly higher levels with 79% CSR

Model	CSR(%) ↑	AMCL ↑	MCL ↑	RF ↓	WS ↑	DUA(%) ↑	ICRA(%) ↑
Open-source models							
LLaVA-Onevision	7	5.20	4	3.14	30.85	67.1	15.2
Video-LLaVA	10	5.15	5	3.13	32.94	68.6	20.1
Phi4-multimodal-instruct	13	5.33	4	3.01	33.78	70.1	21.4
Qwen2.5-VL-7B	16	5.61	9	2.68	35.42	71.0	21.8
InternVL3-8B	19	5.59	8	2.87	35.26	69.2	23.1
Gemma3-12b-it	21	5.53	11	2.81	36.22	72.9	24.5
InternVL3-38B	24	5.75	13	2.57	36.89	75.3	25.1
Qwen2.5-VL-72B	26	5.89	17	2.47	37.69	76.1	25.2
Gemma3-27b-it	29	5.94	20	2.42	37.64	77.7	26.3
Proprietary models							
Gemini-2.0-Flash	31	6.04	21	2.45	39.60	79.4	27.1
Claude-3.5-Sonnet-20241022	35	5.87	23	2.37	38.58	80.9	28.5
GPT-4o-2024-11-20	39	5.94	23	2.17	38.88	80.0	29.7
Gemini-2.0-Flash-thinking	41	6.15	25	2.15	40.65	81.1	30.2
GPT-4.1-2025-04-14	42	6.63	26	1.85	45.59	82.8	32.3
Gemini-2.5-Flash	48	6.90	27	**1.68**	47.63	84.6	36.2
o4-mini-2025-04-16	**51**	**7.19**	**30**	1.69	**55.06**	**85.2**	**39.8**
Best Performance of Models	51	7.19	30	1.68	55.06	85.2	39.8
Human	79	8.03	46	0.74	62.39	92.0	76.8
Maximum	100	8.76	51	0	68.76	100.0	100.0

Table 3: Performance comparison of open-source and proprietary models on CausalStep using seven diagnostic metrics. CSR (%): chain success rate (↑), AMCL: average maximum chain length per video (↑), MCL: global maximum chain length (↑), RF: average restart frequency (↓), WS: weighted score (↑), DUA (%): descriptive understanding accuracy (↑), and ICRA (%): isolated causal reasoning accuracy (↑). The last three rows report the best model performance, average human performance, and theoretical maximum (upper bound for perfect chains).

and an MCL of 46. This disparity is further emphasized by the AMCL and WS, where human performance consistently exceeds that of all models. These results collectively indicate that existing models largely struggle to maintain long, uninterrupted reasoning chains, exhibiting a propensity for frequent interruptions as evidenced by their comparatively higher RF values. This suggests a primary limitation in sustaining deep, multi-step causal reasoning over extended video sequences.

Open-source vs. Proprietary Models. A clear stratification in capabilities is observed between open-source and proprietary models. Proprietary models, exemplified by o4-mini and the Gemini series, consistently demonstrate superior performance across all metrics. For instance, o4-mini surpasses the best open-source model, Gemma-3-27b-it, by a notable 22 points in CSR and 10 points in MCL. This divergence highlights the impact of more extensive resources, larger and more diverse training data, and sophisticated architectural designs prevalent in proprietary systems. However, despite their lead, even the most advanced proprietary models remain considerably behind human-level performance, signaling that significant advancements are still required to bridge this fundamental reasoning gap.

Single-step Understanding vs. Stepwise Reasoning. A granular examination of the single-step metrics—DUA and ICRA—sheds light on distinct challenges. While top models achieve relatively high DUA (up to 85.2%), indicating competence in isolated perceptual understanding, their ICRA remains markedly lower (best model at 39.8%). This stark contrast, coupled with humans' strong ICRA (76.8%), underscores a critical limitation: current models struggle to perform accurate causal reasoning when presented solely with an isolated segment pair, without the benefit of a preceding, correctly established reasoning chain. This discrepancy validates the necessity of CausalStep's stepwise protocol, which inherently evaluates the ability to build and leverage contextual reasoning through sequential reasoning.

Conclusion

We introduce CausalStep, a diagnostic benchmark for explicit stepwise causal reasoning in videos. By segmenting videos into causally linked units and enforcing a strict stepwise QA protocol, CausalStep enables rigorous evaluation of sequential causal reasoning. Our experiments show a huge gap between current MLLMs and human-level performance, particularly in maintaining long reasoning chains. While proprietary models outperform open-source ones, neither approaches human ability on the challenging task. These findings underscore the value of CausalStep for diagnosing model limitations and highlight the need for more robust and causally aware video reasoning systems.

Acknowledgments

This work is supported by Zhongguancun Academy Project No.C20250204, the National Key R&D Program of China (2024YFA1014003), National Natural Science Foundation of China (92470121, 62402016), and High-performance Computing Platform of Peking University.

References

Abouelenin, A.; Ashfaq, A.; et al. 2025. Phi-4-Mini Technical Report: Compact yet Powerful Multimodal Language Models via Mixture-of-LoRAs. *CoRR*, abs/2503.01743.

Anthropic. 2024. Anthropic: Introducing Claude 3.5 Sonnet.

Cai, M.; Tan, R.; Zhang, J.; Zou, B.; Zhang, K.; Yao, F.; Zhu, F.; Gu, J.; Zhong, Y.; Shang, Y.; et al. 2024. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*.

Cao, P.; Men, T.; Liu, W.; Zhang, J.; Li, X.; Lin, X.; Sui, D.; Cao, Y.; Liu, K.; and Zhao, J. 2025. Large language models for planning: A comprehensive and systematic survey. *arXiv preprint arXiv:2505.19683*.

Chen, Z.; Wu, J.; Wang, W.; Su, W.; Chen, G.; Xing, S.; Zhong, M.; Zhang, Q.; Zhu, X.; Lu, L.; et al. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 24185–24198.

Cheng, J.; Ge, Y.; Wang, T.; Ge, Y.; Liao, J.; and Shan, Y. 2025. Video-Holmes: Can MLLM Think Like Holmes for Complex Video Reasoning? *arXiv preprint arXiv:2505.21374*.

Fei, H.; Wu, S.; Ji, W.; Zhang, H.; Zhang, M.; Lee, M.-L.; and Hsu, W. 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition. *arXiv preprint arXiv:2501.03230*.

Fu, C.; Dai, Y.; Luo, Y.; Li, L.; Ren, S.; Zhang, R.; Wang, Z.; Zhou, C.; Shen, Y.; Zhang, M.; et al. 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*, 24108–24118.

Google DeepMind. 2025. Gemini 2.5: Our most intelligent AI model.

Hu, K.; Wu, P.; Pu, F.; Xiao, W.; Zhang, Y.; Yue, X.; Li, B.; and Liu, Z. 2025. Video-MMMU: Evaluating Knowledge Acquisition from Multi-Discipline Professional Videos. *arXiv preprint arXiv:2501.13826*.

Hu, S.; Li, X.; Li, X.; Zhang, J.; Wang, Y.; Zhao, X.; and Cheong, K. H. 2024. Can LVLMs Describe Videos like Humans? A Five-in-One Video Annotations Benchmark for Better Human-Machine Comparison. *arXiv preprint arXiv:2410.15270*.

Hu, S.; Zhang, D.; Feng, X.; Li, X.; Zhao, X.; Huang, K.; et al. 2023. A multi-modal global instance tracking benchmark (mgit): Better locating target in complex spatio-temporal and causal relationship. *Advances in Neural Information Processing Systems*, 36: 25007–25030.

Ibrahim, S. W. 2016. A comprehensive review on intelligent surveillance systems. *Communications in science and technology*, 1(1).

Jiang, J. J.; and Li, X. 2024. Look ahead text understanding and llm stitching. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, 751–760.

Kamath, A.; Ferret, J.; et al. 2025. Gemma 3 Technical Report. *CoRR*, abs/2503.19786.

Kong, Y.; and Fu, Y. 2022. Human action recognition and prediction: A survey. *International Journal of Computer Vision*, 130(5): 1366–1401.

Li, B.; Zhang, Y.; Guo, D.; Zhang, R.; Li, F.; Zhang, H.; Zhang, K.; Zhang, P.; Li, Y.; Liu, Z.; et al. 2024a. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*.

Li, K.; Wang, Y.; He, Y.; Li, Y.; Wang, Y.; Liu, Y.; Wang, Z.; Xu, J.; Chen, G.; Luo, P.; et al. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 22195–22206.

Li, X.; Feng, X.; Hu, S.; Wu, M.; Zhang, D.; Zhang, J.; and Huang, K. 2024c. Dtllm-vlt: Diverse text generation for visual language tracking based on llm. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 7283–7292.

Li, X.; Hu, S.; Feng, X.; Zhang, D.; Wu, M.; Zhang, J.; and Huang, K. 2024d. Dtvlt: A multi-modal diverse text benchmark for visual language tracking based on llm. *arXiv preprint arXiv:2410.02492*.

Li, X.; Hu, S.; Feng, X.; Zhang, D.; Wu, M.; Zhang, J.; and Huang, K. 2024e. How texts help? a fine-grained evaluation to reveal the role of language in vision-language tracking. *arXiv preprint arXiv:2411.15600*.

Li, X.; Hu, S.; Feng, X.; Zhang, D.; Wu, M.; Zhang, J.; and Huang, K. 2024f. Visual language tracking with multi-modal interaction: A robust benchmark. *arXiv preprint arXiv:2409.08887*.

Li, X.; Li, X.; Hu, S.; Guo, Y.; and Zhang, W. 2025. VerifyBench: A Systematic Benchmark for Evaluating Reasoning Verifiers Across Domains. *arXiv preprint arXiv:2507.09884*.

Li, Y.; Chen, X.; Hu, B.; Wang, L.; Shi, H.; and Zhang, M. 2024g. Videovista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*.

Lin, B.; Ye, Y.; Zhu, B.; Cui, J.; Ning, M.; Jin, P.; and Yuan, L. 2023. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*.

Liu, Y.; Li, S.; Liu, Y.; Wang, Y.; Ren, S.; Li, L.; Chen, S.; Sun, X.; and Hou, L. 2024a. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*.

Liu, Y.; Ma, Z.; Qi, Z.; Wu, Y.; Shan, Y.; and Chen, C. W. 2024b. Et bench: Towards open-ended event-level video-language understanding. *Advances in Neural Information Processing Systems*, 37: 32076–32110.

MacKenzie, I. S. 2024. Human-computer interaction: An empirical research perspective.

OpenAI. 2024a. GPT-4o mini: advancing cost-efficient intelligence.

OpenAI. 2024b. OpenAI: Hello GPT-4o.

OpenAI. 2025. Introducing GPT-4.1 in the API. https://openai.com/index/gpt-4-1/.

OpenAI. 2025. OpenAI: Introducing OpenAI o3 and o4-mini.

Qi, Y.; Zhao, Y.; Zeng, Y.; Bao, X.; Huang, W.; Chen, L.; Chen, Z.; Zhao, J.; Qi, Z.; and Zhao, F. 2025. Vcr-bench: A comprehensive evaluation framework for video chain-of-thought reasoning. *arXiv preprint arXiv:2504.07956*.

Reid, M.; Savinov, N.; Teplyashin, D.; Lepikhin, D.; Lillicrap, T.; Alayrac, J.-b.; Soricut, R.; Lazaridou, A.; Firat, O.; Schrittwieser, J.; et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*.

Roy, N.; Posner, I.; Barfoot, T.; Beaudoin, P.; Bengio, Y.; Bohg, J.; Brock, O.; Depatie, I.; Fox, D.; Koditschek, D.; et al. 2021. From machine learning to robotics: Challenges and opportunities for embodied intelligence. *arXiv preprint arXiv:2110.15245*.

Tang, Y.; Bi, J.; Xu, S.; Song, L.; Liang, S.; Wang, T.; Zhang, D.; An, J.; Lin, J.; Zhu, R.; et al. 2025. Video understanding with large language models: A survey. *IEEE Transactions on Circuits and Systems for Video Technology*.

Wang, A.; Wu, B.; Chen, S.; Chen, Z.; Guan, H.; Lee, W.-N.; Li, L. E.; and Gan, C. 2024a. Sok-bench: A situated video reasoning benchmark with aligned open-world knowledge. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 13384–13394.

Wang, W.; He, Z.; Hong, W.; Cheng, Y.; Zhang, X.; Qi, J.; Gu, X.; Huang, S.; Xu, B.; Dong, Y.; et al. 2024b. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*.

Wang, Y.; Wu, S.; Zhang, Y.; Yan, S.; Liu, Z.; Luo, J.; and Fei, H. 2025. Multimodal chain-of-thought reasoning: A comprehensive survey. *arXiv preprint arXiv:2503.12605*.

Wei, H.; Yuan, Y.; Lan, X.; Ke, W.; and Ma, L. 2025. Instructionbench: An instructional video understanding benchmark. *arXiv preprint arXiv:2504.05040*.

Wu, B.; Yu, S.; Chen, Z.; Tenenbaum, J. B.; and Gan, C. 2024a. Star: A benchmark for situated reasoning in real-world videos. *arXiv preprint arXiv:2405.09711*.

Wu, H.; Li, D.; Chen, B.; and Li, J. 2024b. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37: 28828–28857.

Xu, D.; Zhao, Z.; Xiao, J.; Wu, F.; Zhang, H.; He, X.; and Zhuang, Y. 2017. Video question answering via gradually refined attention over appearance and motion. In *Proceedings of the 25th ACM international conference on Multimedia*, 1645–1653.

Yang, A.; Yang, B.; et al. 2024. Qwen2.5 Technical Report. *CoRR*, abs/2412.15115.

Yang, J.; Yang, S.; Gupta, A. W.; Han, R.; Fei-Fei, L.; and Xie, S. 2025. Thinking in space: How multimodal large language models see, remember, and recall spaces. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 10632–10643.

Zhang, Y.; Wu, J.; Li, W.; Li, B.; Ma, Z.; Liu, Z.; and Li, C. 2024. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*.

Zhang, Y.; Zhang, R.; Gu, J.; Zhou, Y.; Lipka, N.; Yang, D.; and Sun, T. 2023. Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint arXiv:2306.17107*.

Zhao, Y.; Zhang, H.; Xie, L.; Hu, T.; Gan, G.; Long, Y.; Hu, Z.; Chen, W.; Li, C.; Xu, Z.; et al. 2025. Mmvu: Measuring expert-level multi-discipline video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, 8475–8489.

Zhou, J.; Shu, Y.; Zhao, B.; Wu, B.; Xiao, S.; Yang, X.; Xiong, Y.; Zhang, B.; Huang, T.; and Liu, Z. 2024. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*.

Zhu, J.; Wang, W.; Chen, Z.; Liu, Z.; Ye, S.; Gu, L.; Duan, Y.; Tian, H.; Su, W.; Shao, J.; et al. 2025a. InternVL3: Exploring Advanced Training and Test-Time Recipes for Open-Source Multimodal Models. *arXiv preprint arXiv:2504.10479*.

Zhu, K.; Jin, Z.; Yuan, H.; Li, J.; Tu, S.; Cao, P.; Chen, Y.; Liu, K.; and Zhao, J. 2025b. MMR-V: What's Left Unsaid? A Benchmark for Multimodal Deep Reasoning in Videos. *arXiv preprint arXiv:2506.04141*.