VCRBench: Exploring Long-form Causal Reasoning Capabilities of Large Video Language Models

Pritam Sarkar

Queen's University, Canada and Vector Institute pritam.sarkar@queensu.ca

Ali Etemad

Queen's University, Canada ali.etemad@queensu.ca







Abstract

Despite recent advances in video understanding, the capabilities of Large Video Language Models (LVLMs) to perform video-based causal reasoning remains underexplored, largely due to the absence of relevant and dedicated benchmarks for evaluating *causal reasoning* in visually grounded and goal-driven settings. To fill this gap, we introduce a novel benchmark named Video-based long-form Causal Reasoning (VCRBench). We create VCRBench using procedural videos of simple everyday activities, where the steps are deliberately shuffled with each clip capturing a key causal event, to test whether LVLMs can identify, reason about, and correctly sequence the events needed to accomplish a specific goal. Moreover, the benchmark is carefully designed to prevent LVLMs from exploiting linguistic shortcuts, as seen in multiple-choice or binary OA formats, while also avoiding the challenges associated with evaluating open-ended QA. Our evaluation of state-ofthe-art LVLMs on VCRBench suggests that these models struggle with video-based long-form causal reasoning, primarily due to their difficulty in modeling long-range causal dependencies directly from visual observations. As a simple step toward enabling such capabilities, we propose Recognition-Reasoning Decomposition (RRD), a modular approach that breaks video-based causal reasoning into two subtasks of video recognition and causal reasoning. Our experiments on VCRBench show that RRD significantly boosts accuracy on VCRBench, with gains of up to 25.2%. Finally, our thorough analysis reveals interesting insights into the reasoning capabilities of LVLMs, for instance, that they primarily rely on their language knowledge when tackling complex video-based long-form causal reasoning tasks.

1 Introduction

Long-form causal reasoning in video involves structured and goal-directed analysis of sequences of visual events. Such capabilities are essential for real-world applications such as household and industrial robotics [1, 2], embodied AI agents [3, 4, 5], spatial intelligence systems [6, 7, 8], and assistive technologies [9, 10], all of which rely on reasoning about causally dependent visual events. While recent advances in vision-language modeling [11, 12, 13] have led to the development of powerful large video language models (LVLMs) [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28], their ability to perform long-form causal-reasoning based on visual observations remains largely underexplored. This is in part due to the lack of benchmarks specifically designed to evaluate causal reasoning in visually-grounded goal-driven settings. In this work, we take a step toward filling

Preprint. Under review.

The given video consists of multiple clips, each showing a different segment needed to complete the task: Make Lemonade. These clips are randomly shuffled, and your job is to arrange them in the correct order to complete the task: Make Lemonade. The clip numbers are mentioned at the beginning of each clip as Clip i, Clip 2, and so on. In order to solve this task, first, you should identify the activity that is performed in each clip, and then use your reasoning and common sense to arrange these clips to successfully complete the task. The final output should be in this format:

Correct order: <mention</pre> the Clip numbers separated by a comma>



Figure 1: **Example question and video.** We present an example of video-based long-form causal reasoning task from VCRBench. *The correct order is: Clip 1: Cut lemon into slices, Clip 5: Squeeze lemon into the pitcher, Clip 4: Pour lemon juice and water into the pitcher, Clip 3: Stir the lemonade mixture, Clip 2: Pour lemonade into a glass.*

this gap by systematically evaluating the video-based causal reasoning capabilities of state-of-the-art LVLMs through a new benchmark. Building on this, we also design a simple modular approach to enhance LVLM performance on video-based long-form causal reasoning tasks.

To study the video-based causal reasoning capabilities of LVLMs, we introduce Video-based longform Causal Reasoning Benchmark (VCRBench), an evaluation benchmark consisting of procedural videos depicting everyday human activities, such as making lemonade or grilling steak (see Figure 1 for an example). VCRBench is designed to evaluate whether LVLMs can identify and reason about visual events with long-form causal dependencies towards a specific goal. Specifically, when presented with a shuffled sequence of video clips each showing a key action, the model must first interpret the actions in each clip and then arrange them in the correct chronological order based on their causal dependencies to complete the procedure. Unlike prior benchmarks [29, 3, 30, 31], VCRBench explicitly tests multi-step causal reasoning and fine-grained spatio-temporal understanding without allowing linguistic shortcuts common in multiple-choice or binary QA formats. For instance, in the lemonade-making example (Figure 1), the model must first distinguish between fine-grained actions such as cutting and squeezing a lemon, and subsequently infer that cutting the lemon, squeezing it, and pouring the juice into a pitcher should occur in the correct causal sequence. Our evaluation across both open- and closed-source models shows that current LVLMs struggle with video-based long-form causal reasoning as most perform at or below random guess, and even the best models fall short of human performance by nearly 40%. Further analysis reveals that while these models can often recognize individual actions, they frequently fail to establish meaningful connections across a sequence of visual events, lacking an understanding of causal dependencies based on visual observations.

To improve the long-form causal reasoning capabilities of LVLMs we introduce *Recognition-Reasoning Decom***position** (RRD), a simple modular approach designed to enhance the video-based reasoning abilities of LVLMs. RRD breaks down video-based long-form causal reasoning into two interconnected sub-tasks: (i) video recognition and (ii) causal reasoning. This decomposition simplifies the overall task by first directing the model's focus toward recognizing visual events, and then reasoning about their relationships to infer the correct causal order. RRD leads to significant gains, improving accuracies by up to 25.2% on VCRBench. Notably, Qwen2.5-VL_{72B}-Instruct, when equipped with RRD, surpasses Gemini-1.5-Pro and achieves a performance comparable to the current topperforming reasoning-specialized model, Gemini-2-Flash-Thinking (see Figure 2).

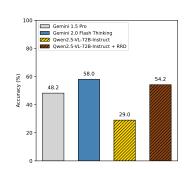


Figure 2: **Impact of RRD.** Qwen2.5-VL-Instruct_{72B} with RRD outperforms Gemini-1.5-Pro and achieve comparable performance to Gemini-2-Flash-Thinking.

In summary, our contributions are as follows:

• We introduce **VCRBench**, a novel benchmark designed to evaluate LVLMs on video-based long-form causal reasoning. To the best of our knowledge, this is the first video evaluation benchmark to

- study multi-step causal reasoning capabilities of LVLMs. Our analysis on various state-of-the-art LVLMs reveals that current LVLMs struggle with long-form causal reasoning due to their inability of meaningfully connect a series of visual events toward a goal.
- To improve the performance of open-source LVLMs on VCRBench, we introduce **RRD**, which decomposes video-based causal reasoning into two related sub-tasks video recognition and causal reasoning. This simple modular approach allows LVLMs to focus on one type of task at a time, first recognition, then reasoning, which results in notable performance gains of up to 25.2%.

2 Background

Large Video Language Models (LVLMs). LVLMs typically consist of a vision encoder, a Large Language Model (LLM), and a cross-modal adapter that bridges visual and textual modalities [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28]. While this high-level structure is common, recent work has introduced considerable architectural variations. These include extending the LLM's context window for long sequences [32, 21], dynamic projection techniques that drop redundant frames based on visual similarity [14, 33], and query-based projectors that selectively attend to relevant visual content [16, 13, 34, 35]. In addition to architectural differences, LVLMs vary in their use of vision encoders, ranging from single to multi-encoder setups (e.g., video + image) [36], and from vision-only to vision-language pretrained models [14]. Training strategies also differ, with some models trained in a single stage, and others using multi-stage pipelines that separate large-scale pretraining (for modality alignment) from instruction tuning or reasoning-specialized post-training [37, 38]. To ensure a comprehensive evaluation of LVLM capabilities across diverse architectural and pretraining paradigms, we have carefully selected models that represent a broad spectrum within these categories for evaluating on VCRBench.

Video evaluation benchmarks. Numerous evaluation benchmarks exist for video understanding tasks, focusing on areas such as information retrieval-based question answering (e.g., ActivityNetQA [39], MSRVTTQA [40], MSVDQA [40], NextQA [41], TGIFQA [42]), comprehensive video understanding (e.g., MVBench [16], TVBench [43], VideoMME [44]), fine-grained temporal understanding (e.g., TVBench [43], TempCompass [45], TemporalBench [46]), long-video understanding (e.g., MLVU [47], LongVideoBench [48]), egocentric video understanding (e.g., Egoschema [49]), and video hallucination (e.g., VideoHallucer [50], HallusionBench [51]), among others. There also exist a few benchmarks focused on video-based reasoning, such as SOK-Bench [52], MMWorld [53], and VILMA [54]. However, a significant gap remains in the evaluation of video-based causal reasoning tasks. While some benchmarks address intent (e.g., IntentQA [29]), causal question answering (e.g., Causal-VidQA [31]), or goal-oriented question answering (e.g., EgoPlan-Bench [3], ReXTime [30]), they do not adequately assess the video-based long-form or multi-step causal reasoning capabilities of LVLMs. In this work, we address the critical area of long-form causal reasoning, which refers to reasoning about visual events with multiple or interconnected causal dependencies.

Reasoning methods. Chain-of-Thought prompting has emerged as a powerful technique to improve reasoning in LLMs and LVLMs by encouraging intermediate step-by-step derivations rather than direct answer prediction [55]. This paradigm has been further strengthened by post-training alignment techniques such as Reinforcement Learning with Human Feedback (RLHF) [56, 57], which optimize models to generate more helpful and aligned responses. More recent methods like DeepSeek's R1 [58] also build on such alignment strategies to enhance reasoning quality. In parallel, a growing body of work explores inference-time techniques to boost performance without the necessity of additional training. These include majority voting or self-consistency sampling [59], which aggregate multiple generated responses for robustness, best-of-N sampling [60, 61], which selects the highest-quality sample from multiple candidates, and decomposed prompting [62], which breaks complex reasoning tasks into simpler sub-tasks. Our proposed approach, RRD, is motivated by decomposed prompting where complex video-based reasoning tasks are systematically divided into several sub-tasks.

3 Video-based long-form Causal Reasoning Benchmark (VCRBench)

3.1 Construction of VCRBench

We construct VCRBench by curating a set of everyday procedures that require no specialized knowledge and are commonly encountered in daily life, such as grilling steak, making lemonade, or



Figure 3: **Overview of video construction. Step 1:** Given a complete video, key procedural steps are identified based on human-annotated timestamps. **Step 2:** We keep the key events and discard those that do not depict visual events directly associated with the goal, such as talking or narrating in this example of grilling steak. **Step 3:** Each key event is shuffled across time and assigned a clip number. These clips are then merged together to form the final test sample.

The given video consists of multiple short clips, each showing a different segment needed to complete the task: {name of the procedure}.

These clips are randomly shuffled, and your job is to arrange them in the correct order to complete the task: {name of the procedure}.

The clip numbers are mentioned at the beginning of each clip as Clip 1, Clip 2, and so on.

In order to solve this task, first, you should identify the activity that is performed in each clip, and then use your reasoning and common sense to arrange these clips to successfully complete the task.

The final output should be in this format:

Correct order: <mention the Clip numbers separated by a comma>

Figure 4: The **question template** used in VCRBench.

preparing pancakes (see Appendix B for the full list). For each procedure, we source instructional videos from the CrossTask dataset [63], which contains YouTube videos with human annotated timestamps of key events. Below, we outline the three-stage process for preparing videos and questions in VCRBench.

Preparing the videos. Our video construction pipeline consists of the following steps:

- **Step 1.** We begin with a complete procedural video and use the provided human-annotated timestamps to segment it into short clips, each corresponding to a specific procedural step (e.g., *seasoning steak* for the procedure *grill steak*).
- Step 2. Using WikiHow as a reference (https://www.wikihow.com/), we identify the core steps necessary for the procedure. We group consecutive steps that have no causal dependencies. Moreover, we remove irrelevant segments that do not contribute to the main task, ensuring that all selected clips exhibit causal dependencies. At this stage, the resulting set of clips must follow a meaningful chronological order for successful completion of the procedure. This step is manually curated by human annotators to ensure accurate assessment.
- Step 3. The selected clips are then randomly shuffled, with the constraint that the original ascending order is not retained. The shuffled clips are concatenated into a single video, with blank frames inserted between them for visual separation. The blank frames preceding the clips labeled chronologically to clearly distinguish the individual steps. The resulting video serves as the input to the LVLM, which is tasked with identifying the correct order of the procedural steps.

An overview of this construction process is illustrated in Figure 3.

Preparing the questions. A key challenge in evaluating LVLMs is designing a reliable evaluation protocol to correctly asses true visual understanding. Most existing video benchmarks [16, 43, 50, 44, 48, 45, 39, 40, 41, 42, 49, 51, 3, 29] rely on multiple-choice or binary question-answering formats, to streamline their automated evaluation. However, these formats can be exploited through linguistic cues in the provided response choices, without requiring true visual understanding. Open-ended question answering offers a more rigorous probe of visual reasoning, but introduces evaluation ambiguity which often necessitates the use of an external LLM (e.g., GPT-4 [64]) as a judge for automated evaluation, a strategy proven to be unreliable and ambiguous in prior work [43]. VCRBench addresses

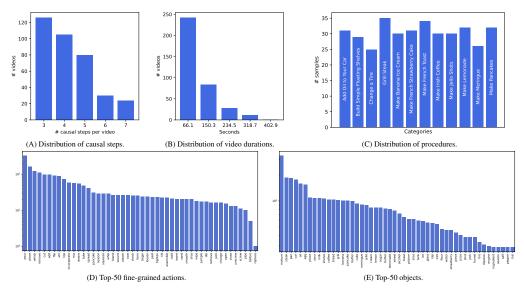


Figure 5: Key statistics of our VCRBench.

these limitations by framing causal reasoning as a sequence ordering task. This setup avoids the use of linguistic cues in predefined options, yet yields deterministic ground truth answers. As a result, it enables accurate, objective evaluation while still challenging LVLMs to perform fine-grained visual and causal reasoning. The question template is mentioned in Figure 4.

Statistics. VCRBench comprises 365 videos and questions across 12 categories of procedures covering diverse fine-grained actions and object interactions. The videos are 30 to 445 seconds long with an average duration of 107 seconds and a total duration of 10 hours. To keep the difficulty reasonable, we include videos requiring only 3 to 7 causally dependent steps with an average of 4.2 steps per task. Additional key statistics are provided in Figure 5.

3.2 Evaluation Metrics

We measure the performance of LVLMs on VCRBench with two metrics: overall accuracy and step accuracy [65]. Overall accuracy (also referred to simply as Accuracy) indicates predictions that exactly match the ground truth, whereas step accuracy compares predicted and ground truth actions step by step. Assume, $(q,v) \in \mathcal{D}$, where q is a question related to a video v sampled from a validation set \mathcal{D} . Let π be an LVLM and $\mathbb{1}(\cdot)$ be the indicator function of correct prediction. The mathematical expressions of our evaluation metrics are as follows:

$$\operatorname{Accuracy} = \frac{\sum_{(q,v) \in \mathcal{D}} \mathbbm{1}\left(\pi(q,v)\right)}{|\mathcal{D}|}, \qquad \operatorname{Step Accuracy} = \frac{\sum_{(q,v) \in \mathcal{D}} \frac{\sum_{s} \mathbbm{1}\left(\pi(q,v)\right)}{s}}{|\mathcal{D}|},$$

where s denotes the total number of steps to a procedure.

4 Benchmarking Results

4.1 Setup

We examine over 20 recent and popular LVLMs, including both closed and open-source models. These models exhibit significant variations in several key aspects: LLM architectures (LLaMA [66, 67, 68], Mistral [69], and Qwen [70, 71]) with sizes from 1B to 78B parameters for open-source LVLMs; cross-modal adapters (QFormer [72], MLP projector [73], and spatio-temporal compressor [14, 18]); vision encoders with single or dual configurations (CLIP [11], SigLIP [12], DINO [74], and UMT [75]); training methodologies (single-stage or multi-stage) including alignment finetuning for improved reasoning ([28, 76]); and visual frame processing capabilities ranging from 8 (NVILA [77]) to over 500 frames (LongVU [14], Qwen2.5-VL [28]). We follow the recommended generation configurations, such as temperature, system prompt, number of frames, and other key parameters, for each respective LVLM. For reference, we also benchmark human performance on VCRBench.

Table 1: **Results on VCRBench.** Most open-source LVLMs perform at or below random guess, and even the best LVLM falls significantly short of human performance. We fade numbers that fall below the random guess baseline.

Models	# Frame	s Acc. _(†) S	Step Acc. (†)
Random Guess		7.8	24.1
InternVL2.5 _{1B} [26]	64	1.4	10.3
InternVL2.5 _{2B} [26]	64	6.3	16.2
LongVU _{3B} [14]	1fps	0.0	7.0
InternVL2.5 _{4B} [26]	64	1.6	9.5
VideoChat2 _{7B} [16]	16	0.3	5.8
InternVL2.5 _{8B} [26]	64	2.7	11.1
LLaVA-NeXT-Video _{7B} [22]	64	0.0	17.4
MiniCPM-o-V 2.67B [78]	64	2.5	11.0
Qwen2.5-VL-Instruct _{7B} [28]	1fps	7.1	20.9
VideoLLaMA37B [19]	128	1.6	13.1
LongVILA7B [21]	128	0.3	1.1
LongVU _{7B} [14]	1fps	0.0	2.4
NVILA _{15B} [77]	8	0.6	3.6
InternVL2.5 _{26B} [26]	64	2.7	13.7
InternVL2.5 _{38B} [26]	64	11.0	27.4
LLaVA-NeXT-Video _{72B} [22]	32	5.2	18.6
Qwen2.5-VL-Instruct _{72B} [28]	1fps	29.0	44.0
InternVL2.5 _{78B} [26]	64	14.5	34.0
GPT4o [79]	32	29.0	36.6
Gemini-1.5-Pro [24]	1fps	48.2	65.3
Gemini-2.0-Flash-Thinking [24]	1fps	58.0	67.7
Human		96.4	98.3

# Response from Correct Order:	LongVILA Clip 1, Clip 2,, Clip 14,
# Response from Correct Order:	LongVU 1, 2, 3, 4, 5,, 29, 30,
# Response from Correct Order:	LLaVA-Next-Qwen Clip 1, Clip 2, Clip 3, Clip 4

Figure 6: **Failure examples.** Several opensource LVLMs merely list consecutive numbers as the predicted order, exhibiting inability to make a meaningful attempt in VCRBench tasks.

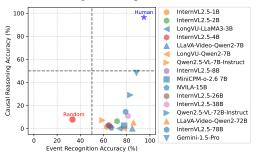


Figure 7: **Reasoning vs. Recognition.** LVLMs can identify individual events but fail to connect them toward a specific goal in causal reasoning.

4.2 Results and Findings

Here, we discuss our key observations regarding the performance of LVLMs on VCRBench, based on our detailed quantitative and qualitative analysis.

VCRBench tasks are unambiguous to human evaluators. As shown in Table 1, human participants achieve an accuracy of 96.4% on VCRBench. This high performance indicates that the video-based long-form causal reasoning tasks are intuitive and unambiguous to humans. Further details on the human evaluation setup are provided in Appendix B.

LVLMs lack video-based long-form causal reasoning abilities. As shown in Table 1, most open-source LVLMs perform worse than random guessing, with the exception of InternVL_{38B}, Qwen2.5-VL-Instruct_{72B}, and InternVL_{78B}. Several open-source LVLMs (e.g., LongVILA, LLaVA-NeXT-Video, LongVU) exhibit a tendency to output a sequence of consecutive numbers, up to their maximum generation length, as the presumed correct order, see examples in Figure 6. This suggests that these models have not developed a robust notion of attempting video-based causal reasoning tasks in VCRBench, and instead default to token-level statistical regularities when uncertain. Surprisingly, even open-source models built for improved reasoning, such as MiniCPM-o [78], underperform on VCRBench, suggesting limited video-based causal reasoning abilities. Among open-source models, Qwen2.5-VL-Instruct_{72B} performs best, though it still lags significantly behind the best closed-source model, Gemini-2-Flash-Thinking. Interestingly, GPT-40 performs worst among the closed-source models, likely due to its limited capacity for long visual inputs, achieving a performance similar to Qwen2.5-VL-Instruct_{72B}. Overall, even the best-performing model, Gemini-2-Flash-Thinking, falls substantially short of human-level performance (58.0% vs. 96.4%). We present sample responses from top-performing LVLMs in Figure 8 and additional results in Appendix A.

Recognizing events is not enough: LVLMs lack associative understanding of visual events. To better understand the limitations of current LVLMs, we conduct additional experiments that evaluate their ability to recognize intermediate events, which is a prerequisite for long-form causal reasoning. We design a multiple-choice question-answering (MCQ) evaluation where each sample presents a clip of an intermediate step along with three answer options: one correct and two incorrect. We derive the incorrect options from the other steps within the procedure. Figure 7 plots video recognition accuracy against causal reasoning accuracy to analyze their relationship. As shown in Figure 7, most LVLMs achieve decent accuracy ($\sim 80\%$) on event recognition, while performing poorly in long-form causal reasoning tasks. This suggests that while LVLMs can recognize individual visual events when

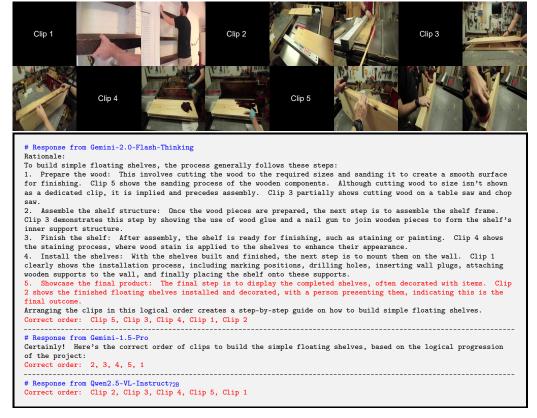


Figure 8: **Failure examples.** We observe that Gemini-2.0-Flash-Thinking generates a response with a detailed rationale explaining how it arrives at the final answer, unlike Gemini-1.5-Pro and Qwen2.5-VL-Instruct, which directly provide the final answer. Based on its detailed response, Gemini-2.0-Flash-Thinking correctly interprets most actions except for Clip 2 (highlighted in red). However, it entirely fails to arrange the identified events according to their causal dependencies. Additionally, both Gemini-1.5-Pro and Qwen2.5-VL-Instruct_{72B} make the same mistake: they fail to recognize the causal link between steps 4 and 5, i.e., *the shelves must be sanded before being painted*. The correct order is 2, 3, 5, 4, 1.

provided answer choices, they struggle with video recognition in an open-ended setup and fail to connect these events meaningfully toward a goal. Our results further suggest that recognizing events is necessary but not sufficient for long-form causal reasoning. For example, although LLaVA-Video-Qwen2_{72B} performs better than Qwen2.5-VL-Instruct_{72B} on the video recognition task, it completely fails at causal reasoning.

5 A Simple Step Towards Improving Video-based Causal Reasoning

5.1 Recognition-Reasoning Decomposition

Humans excel at reasoning by decomposing complex tasks into a series of sub-tasks, addressing each in a sequential manner, and leveraging the intermediate results to arrive at a final conclusion. Inspired by this cognitive problem-solving strategy, we propose a modular approach that explicitly decomposes video-based causal reasoning tasks into two distinct, yet interdependent, sub-problems: (i) video recognition, which aims to extract salient events from the visual input, and (ii) causal reasoning, which involves inferring the causal relationships between these identified events. This decomposition allows the LVLM to focus on one type of task at a time. We refer to this approach as Recognition-Reasoning Decomposition (RRD). The details of the sub-tasks are described as follows:

• **Video recognition.** As the first stage of our approach, we instruct the LVLM to obtain the descriptions of fine-grained actions/events for each clip of the video using the following prompt: Provide a one-sentence description indicating the key and fine-grained

actions or events for each clip. Please respond in this format:

Clip 1: <Write one sentence description>

Clip 2: <Write one sentence description>.

This allows the LVLM to strictly focus on the actions and events without necessarily considering the causal relationships among clips, enabling explicit focus on and localized analysis of each clip.

• Causal reasoning. The next stage of RRD involves arranging the identified events from the video recognition step based on their causal relationships to complete the procedure. Note that the clips are shuffled, and thus, so are the identified events. To this end, we instruct the LVLM to identify the correct order of the events identified in the previous stage, using the following prompt:

The following randomly shuffled steps are needed to complete the task: {name of the procedure}.

Use your reasoning and common sense to arrange these steps to successfully complete the task.

{clip descriptions}

This process enables the LVLM to leverage its language capabilities for reasoning tasks.

5.2 Experiments and Results

To test RRD on our proposed VCRBench, we use the top-performing open-source LVLMs (based on their performance on VCRBench in Table 1), i.e., InternVL2.5 and Qwen2.5-VL-Instruct. Specifically, we use both the 7B and 72B variants of Qwen2.5-VL-Instruct and 38B and 78B variants of InternVL2.5. We follow the recommended inference setup of these LVLMs and use 64 frames for InternVL2.5 and sample frames at 1 FPS for Qwen2.5-VL-Instruct, similar to Table 1. Following we provide our findings regarding the behavior of RRD on VCRBench, along with detailed experimental results and analysis.

RRD significantly improves video-based long-form causal reasoning capabilities of LVLMs. The results in Table 2 demonstrate that our proposed RRD significantly enhances the video-based causal reasoning capabilities of LVLMs. The benefits of RRD are consistent across different model sizes (from 7B to 78B) and across both weaker to stronger LVLMs. For instance, Qwen2.5-VL-Instruct7B, which initially performed close to random guess on VCRBench, achieves a 15.3% accuracy gain when equipped with RRD. Similarly, the 38B and 78B variants of InternVL2.5 show improvements of 12.6% and 13.7%, respectively. Moreover, the top-performing open-source LVLM Qwen2.5-VL-Instruct72B sees a substantial improvement of 20.8% in accuracy when equipped with RRD. Note that

Table 2: **Impact of RRD.** Our proposed task decomposition significantly enhances the long-form causal reasoning capabilities of LVLMs, yielding accuracy improvements of between 12.6% and 20.9%.

Models	Acc. _(↑)	Step Acc. (↑)
Qwen2.5-VL-Instruct _{7B} [28] + RRD (Ours)	$7.1 \\ 22.5_{\uparrow 15.4}$	$20.9 \\ 37.3_{\uparrow 16.4}$
InternVL2.5 _{38B} [26] + RRD (Ours)	$11.0 \\ 23.6_{\uparrow 12.6}$	$27.4 \\ 34.3_{\uparrow 6.9}$
Qwen2.5-VL-Instruct _{72B} [28] + RRD (Ours)	$29.0 \\ 49.9_{\uparrow 20.9}$	43.0 63.4 _{↑20.4}
InternVL2.5 _{78B} [26] + RRD (Ours)	$14.5 \\ 28.2_{\uparrow 13.7}$	34.0 $43.5_{\uparrow 9.5}$

RRD improves the performance of LVLMs that rely on a fixed number of visual inputs, such as InternVL2.5, as well as models that accept a varying number of frames, such as Qwen2.5-VL-Instruct.

LVLMs mainly depend on their language knowledge for complex reasoning while including vision may hinder performance. The results presented in Table 3 suggest that LVLMs mainly rely on their language abilities when solving complex reasoning tasks. Interestingly, we find that incorporating videos in addition to the clip descriptions at the causal reasoning step degrades the accuracy of LVLMs on VCRBench. This performance drop may be due to possible conflicts or misalignment between the visual and linguistic understanding of the model.

Table 3: The effect of incorporating videos at causal reasoning step. The results are based on Qwen2.5-VL-Instruct_{72B}. Here ■ refers to videos and □ refers to generated video descriptions from video recognition step. Using videos at reasoning stage degrade performance.

Video Recognition	Causal Reasoning	Acc. _(↑)	Step Acc. (†)
	à	49.9	63.4
	€ + à	46.6 _{\rightarrow3.3}	62.8 ↓0.6

Sequential recognition improves performance by focusing on one key event at a time, while sequential reasoning degrades overall accuracy due to the need for global context in long-form causal reasoning. Next, we conduct a thorough analysis in the main design of RRD by examining the effect of performing video recognition across all clips (referred to as all-at-once) versus analyzing each clip individually (referred to as sequential). Intuitively, the sequential approach

Table 4: The effect of further decomposing video recognition and causal reasoning steps.

Video Recognition	Causal Reasoning	Acc. _(↑)	Step Acc. (†)
all-at-once	all-at-once	49.9	63.4
all-at-once	sequential	47.4	64.1
sequential	all-at-once	$\bf 54.2$	65.1
sequential	sequential	51.0	66.6

further simplifies the video recognition task and allows the LVLM to focus on localized analysis of one clip at a time. The detailed setup for this experiment is presented in Appendix C. Similarly, for causal reasoning, we explore pairwise causal comparisons in a sequential manner against determining the correct causal order all at once. To perform pairwise comparisons, we adopt a sorting algorithm (i.e., merge sort) that arranges visual events into a causal chain, where each comparison is based on the causality between two events as determined by the LVLM. Upon completion of sorting, the resulting ordered list of events is used as the final prediction. The detailed setup for this experiment is provided in Appendix C. We conduct this experiment utilizing the best open-source LVLM Owen2.5-VL-Instruct_{72B}. The results are presented in Table 4, which reveal the following: (i) performing video recognition sequentially helps LVLMs focus on one key event at a time, leading to improved accuracy in VCRBench; (ii) for causal reasoning, however, the sequential approach does not yield better results. We conjecture that this is due to the long-range dependencies among causal events: presenting all events together allows the LVLM to better capture the global causal structure, whereas pairwise comparisons provide only local views. Although step accuracy, which measures the correctness of individual steps, shows slight improvement, the overall reasoning accuracy is lower in the sequential causal reasoning setup compared to the all-at-once approach.

Table 5: Qwen2.5-VL-Instruct_{72B} equipped with RRD outperforms Gemini-1.5-Pro and achieve a comparable performance to Gemini-2.0-Flash Thinking.

Models		Overall				
	3	4	5	6	7	Acc.
Qwen2.5-VL-Instruct _{72B} [28] + RRD (ours)	$40.5 \\ 64.3_{\uparrow 19.8}$	37.1 73.3 _{\uparrow36.2}	$16.2\\31.2_{\uparrow 15.0}$	$\begin{array}{c} 6.7 \\ 23.3_{\uparrow 16.6} \end{array}$	$4.2 \\ 33.3_{\uparrow 29.1}$	$\begin{array}{c} 29.0 \\ 54.2_{\uparrow 25.2} \end{array}$
Gemini-1.5-Pro [24] Gemini-2.0-Flash-Thinking [24]	60.3 64.8	50.5 66.7	36.2 46.2	43.3 53.3	20.8 29.2	48.2 58.0

RRD effectively enhances video-based causal reasoning across varying number of steps. Table 5 shows that the benefits of RRD are consistent across videos with varying number of steps. Notably, Qwen2.5-VL-Instruct_{72B} equipped with RRD outperforms Gemini-1.5-Pro by 6% and achieves a performance comparable to the reasoning-focused Gemini-2.0-Flash-Thinking. In some setups, it even surpasses Gemini-2.0-Flash-Thinking (see Table 5 for details).

6 Discussions

Summary. In this work, we introduce VCRBench, a novel benchmark designed to evaluate video-based long-form causal reasoning capabilities of LVLMs. Through a comprehensive study of over 20 recent and popular LVLMs, we find that current models consistently struggle with long-form causal reasoning based on visual observations, largely due to a lack of associative understanding of visual events. As an initial step towards enabling such capabilities we introduce RRD, a simple approach that decomposes video-based causal reasoning into video recognition and causal reasoning tasks. RRD significantly improves long-form causal reasoning capabilities of LVLMs, for instance, Qwen2.5-VL-Instruct72B with RRD outperforms Gemini-1.5-Pro and achieves performance comparable to Gemini-2-Flash-Thinking.

Limitations and future work. While RRD effectively improves the video-based causal reasoning capabilities of LVLMs, it relies on explicit human instructions for decomposing complex reasoning tasks. Future work could focus on developing LVLMs that can internalize such modular approach in decomposing complex tasks without explicit human guidance. Furthermore, future work could focus on developing LVLMs capable of performing complex reasoning directly from visual inputs, in contrast to current models that primarily rely on language knowledge for such tasks.

Acknowledgment

We thank Debaditya Shome and Nishq Poorav Desai for their help in building the platform used to collect responses from human evaluators on VCRBench. We also thank members of our lab who volunteered as human evaluators in assessing VCRBench. We thank the Bank of Montreal and Mitacs for funding this research, and the Vector Institute for providing computational resources.

References

- [1] Pierre Sermanet, Tianli Ding, Jeffrey Zhao, Fei Xia, Debidatta Dwibedi, Keerthana Gopalakrishnan, Christine Chan, Gabriel Dulac-Arnold, Sharath Maddineni, Nikhil J Joshi, et al. Robovqa: Multimodal long-horizon reasoning for robotics. In *ICRA*, pages 645–652. IEEE, 2024. 1
- [2] Jinming Li, Yichen Zhu, Zhiyuan Xu, Jindong Gu, Minjie Zhu, Xin Liu, Ning Liu, Yaxin Peng, Feifei Feng, and Jian Tang. Mmro: Are multimodal llms eligible as the brain for in-home robotics? *arXiv preprint arXiv:2406.19693*, 2024. 1
- [3] Yi Chen, Yuying Ge, Yixiao Ge, Mingyu Ding, Bohao Li, Rui Wang, Ruifeng Xu, Ying Shan, and Xihui Liu. Egoplan-bench: Benchmarking multimodal large language models for human-level planning. *arXiv preprint arXiv:2312.06722*, 2023. 1, 2, 3, 4
- [4] Arjun Majumdar, Anurag Ajay, Xiaohan Zhang, Pranav Putta, Sriram Yenamandra, Mikael Henaff, Sneha Silwal, Paul Mcvay, Oleksandr Maksymets, Sergio Arnaud, et al. Openeqa: Embodied question answering in the era of foundation models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16488–16498, 2024. 1
- [5] Liang Chen, Yichi Zhang, Shuhuai Ren, Haozhe Zhao, Zefan Cai, Yuchi Wang, Peiyi Wang, Tianyu Liu, and Baobao Chang. Towards end-to-end embodied decision making via multi-modal large language model: Explorations with gpt4-vision and beyond. *arXiv preprint arXiv:2310.02071*, 2023. 1
- [6] Jie Feng, Jinwei Zeng, Qingyue Long, Hongyi Chen, Jie Zhao, Yanxin Xi, Zhilun Zhou, Yuan Yuan, Shengyuan Wang, Qingbin Zeng, et al. A survey of large language model-powered spatial intelligence across scales: Advances in embodied agents, smart cities, and earth science. *arXiv* preprint arXiv:2504.09848, 2025. 1
- [7] Boyuan Chen, Zhuo Xu, Sean Kirmani, Brain Ichter, Dorsa Sadigh, Leonidas Guibas, and Fei Xia. Spatialvlm: Endowing vision-language models with spatial reasoning capabilities. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14455–14465, 2024. 1
- [8] Wenlong Huang, Chen Wang, Ruohan Zhang, Yunzhu Li, Jiajun Wu, and Li Fei-Fei. Voxposer: Composable 3d value maps for robotic manipulation with language models. *arXiv preprint arXiv:2307.05973*, 2023. 1
- [9] Jie Liu, Wenxuan Wang, Yihang Su, Jingyuan Huan, Wenting Chen, Yudi Zhang, Cheng-Yi Li, Kao-Jung Chang, Xiaohan Xin, Linlin Shen, et al. A spectrum evaluation benchmark for medical multi-modal large language models. *arXiv preprint arXiv:2402.11217*, 2024. 1
- [10] Fan Bai, Yuxin Du, Tiejun Huang, Max Q-H Meng, and Bo Zhao. M3d: Advancing 3d medical image analysis with multi-modal large language models. arXiv preprint arXiv:2404.00578, 2024. 1
- [11] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021. 1, 5
- [12] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11975–11986, 2023. 1, 5
- [13] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Guo Chen, Baoqi Pei, Rongkun Zheng, Jilan Xu, Zun Wang, et al. Internvideo2: Scaling video foundation models for multimodal video understanding. *arXiv preprint arXiv:2403.15377*, 2024. 1, 3

- [14] Xiaoqian Shen, Yunyang Xiong, Changsheng Zhao, Lemeng Wu, Jun Chen, Chenchen Zhu, Zechun Liu, Fanyi Xiao, Balakrishnan Varadarajan, Florian Bordes, et al. Longvu: Spatiotemporal adaptive compression for long video-language understanding. *arXiv preprint* arXiv:2410.17434, 2024. 1, 3, 5, 6, 19
- [15] Mingze Xu, Mingfei Gao, Zhe Gan, Hong-You Chen, Zhengfeng Lai, Haiming Gang, Kai Kang, and Afshin Dehghan. Slowfast-llava: A strong training-free baseline for video large language models. *arXiv preprint arXiv:2407.15841*, 2024. 1, 3
- [16] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, et al. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 1, 3, 4, 6, 19
- [17] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. *arXiv preprint arXiv:2311.17043*, 2023. 1, 3
- [18] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024. 1, 3, 5
- [19] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 1, 3, 6, 19
- [20] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 1, 3
- [21] Yukang Chen, Fuzhao Xue, Dacheng Li, Qinghao Hu, Ligeng Zhu, Xiuyu Li, Yunhao Fang, Haotian Tang, Shang Yang, Zhijian Liu, et al. Longvila: Scaling long-context visual language models for long videos. *arXiv preprint arXiv:2408.10188*, 2024. 1, 3, 6, 19
- [22] Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. Video instruction tuning with synthetic data. *arXiv preprint arXiv:2410.02713*, 2024. 1, 3, 6, 19
- [23] Bin Lin, Bin Zhu, Yang Ye, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. arXiv preprint arXiv:2311.10122, 2023.
 1, 3
- [24] Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean-baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv* preprint arXiv:2403.05530, 2024. 1, 3, 6, 9, 15, 19
- [25] Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, et al. 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*, pages 24185–24198, 2024. 1, 3
- [26] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 1, 3, 6, 8, 15, 19
- [27] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A versatile vision-language model for understanding, localization. *Text Reading, and Beyond*, 2, 2023. 1, 3
- [28] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 1, 3, 5, 6, 8, 9, 15, 19
- [29] Jiapeng Li, Ping Wei, Wenjuan Han, and Lifeng Fan. Intentqa: Context-aware video intent reasoning. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 11963–11974, 2023. 2, 3, 4
- [30] Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Frank Wang. Rextime: A benchmark suite for reasoning-across-time in videos. *Advances in Neural Information Processing Systems*, 37:28662–28673, 2024. 2, 3

- [31] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21273–21282, 2022. 2, 3
- [32] Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. Long context transfer from language to vision. arXiv preprint arXiv:2406.16852, 2024. 3
- [33] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. 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, 2024. 3
- [34] Xinhao Li, Yi Wang, Jiashuo Yu, Xiangyu Zeng, Yuhan Zhu, Haian Huang, Jianfei Gao, Kunchang Li, Yinan He, Chenting Wang, et al. Videochat-flash: Hierarchical compression for long-context video modeling. *arXiv preprint arXiv:2501.00574*, 2024. 3
- [35] Guangzhi Sun, Wenyi Yu, Changli Tang, Xianzhao Chen, Tian Tan, Wei Li, Lu Lu, Zejun Ma, Yuxuan Wang, and Chao Zhang. video-salmonn: Speech-enhanced audio-visual large language models. *arXiv preprint arXiv:2406.15704*, 2024. 3
- [36] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Khan. Videogpt+: Integrating image and video encoders for enhanced video understanding. *arXiv preprint arXiv:2406.09418*, 2024. 3
- [37] Yunlong Tang, Jing Bi, Siting Xu, Luchuan Song, Susan Liang, Teng Wang, Daoan Zhang, Jie An, Jingyang Lin, Rongyi Zhu, et al. Video understanding with large language models: A survey. arXiv preprint arXiv:2312.17432, 2023. 3
- [38] Neelu Madan, Andreas Møgelmose, Rajat Modi, Yogesh S Rawat, and Thomas B Moeslund. Foundation models for video understanding: A survey. *arXiv preprint arXiv:2405.03770*, 2024.
- [39] Zhou Yu, Dejing Xu, Jun Yu, Ting Yu, Zhou Zhao, Yueting Zhuang, and Dacheng Tao. 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, 2019. 3, 4
- [40] Dejing Xu, Zhou Zhao, Jun Xiao, Fei Wu, Hanwang Zhang, Xiangnan He, and Yueting Zhuang. Video question answering via gradually refined attention over appearance and motion. In *ACM Multimedia*, 2017. 3, 4
- [41] Junbin Xiao, Xindi Shang, Angela Yao, and Tat-Seng Chua. 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, 2021. 3, 4
- [42] Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 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, 2017. 3, 4
- [43] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Tybench: Redesigning video-language evaluation. *arXiv preprint arXiv:2410.07752*, 2024. 3, 4
- [44] Chaoyou Fu, Yuhan Dai, Yongdong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, et al. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *arXiv preprint arXiv:2405.21075*, 2024. 3, 4
- [45] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024. 3, 4
- [46] Mu Cai, Reuben Tan, Jianrui Zhang, Bocheng Zou, Kai Zhang, Feng Yao, Fangrui Zhu, Jing Gu, Yiwu Zhong, Yuzhang Shang, et al. Temporalbench: Benchmarking fine-grained temporal understanding for multimodal video models. *arXiv preprint arXiv:2410.10818*, 2024. 3
- [47] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 3

- [48] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2025. 3, 4
- [49] Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Advances in Neural Information Processing Systems*, 36:46212–46244, 2023. 3, 4
- [50] Yuxuan Wang, Yueqian Wang, Dongyan Zhao, Cihang Xie, and Zilong Zheng. Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. *arXiv preprint arXiv:2406.16338*, 2024. 3, 4
- [51] T Guan, F Liu, X Wu, R Xian, Z Li, X Liu, X Wang, L Chen, F Huang, Y Yacoob, et al. Hallusionbench: an advanced diagnostic suite for entangled language hallucination and visual illusion in large vision-language models. arxiv. 10.48550. arXiv preprint arXiv.2310.14566, 2023. 3, 4
- [52] Andong Wang, Bo Wu, Sunli Chen, Zhenfang Chen, Haotian Guan, Wei-Ning Lee, Li Erran Li, and Chuang Gan. 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*, pages 13384–13394, 2024. 3
- [53] Xuehai He, Weixi Feng, Kaizhi Zheng, Yujie Lu, Wanrong Zhu, Jiachen Li, Yue Fan, Jianfeng Wang, Linjie Li, Zhengyuan Yang, et al. Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. *arXiv preprint arXiv:2406.08407*, 2024. 3
- [54] Ilker Kesen, Andrea Pedrotti, Mustafa Dogan, Michele Cafagna, Emre Can Acikgoz, Letitia Parcalabescu, Iacer Calixto, Anette Frank, Albert Gatt, Aykut Erdem, et al. Vilma: A zero-shot benchmark for linguistic and temporal grounding in video-language models. *arXiv preprint arXiv:2311.07022*, 2023. 3
- [55] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35:24824–24837, 2022. 3
- [56] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *NeurIPS*, 30, 2017. 3
- [57] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3
- [58] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 3
- [59] Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022. 3
- [60] Nisan Stiennon, Long Ouyang, Jeffrey Wu, Daniel Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul F Christiano. Learning to summarize with human feedback. Advances in neural information processing systems, 33:3008–3021, 2020.
- [61] Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, et al. Webgpt: Browser-assisted question-answering with human feedback. *arXiv preprint arXiv:2112.09332*, 2021. 3
- [62] Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. Decomposed prompting: A modular approach for solving complex tasks. *arXiv preprint arXiv:2210.02406*, 2022. 3
- [63] Dimitri Zhukov, Jean-Baptiste Alayrac, Ramazan Gokberk Cinbis, David Fouhey, Ivan Laptev, and Josef Sivic. Cross-task weakly supervised learning from instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3537–3545, 2019. 4, 15
- [64] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. arXiv preprint arXiv:2303.08774, 2023. 4

- [65] Chien-Yi Chang, De-An Huang, Danfei Xu, Ehsan Adeli, Li Fei-Fei, and Juan Carlos Niebles. Procedure planning in instructional videos. In *European Conference on Computer Vision*, pages 334–350. Springer, 2020. 5
- [66] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 5
- [67] Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288, 2023. 5
- [68] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 5
- [69] Fengqing Jiang. Identifying and mitigating vulnerabilities in llm-integrated applications. Master's thesis, University of Washington, 2024.
- [70] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023. 5
- [71] An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al. Qwen2. 5 technical report. arXiv preprint arXiv:2412.15115, 2024. 5
- [72] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International conference on machine learning*, pages 19730–19742. PMLR, 2023. 5
- [73] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26296–26306, 2024. 5
- [74] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *arXiv preprint arXiv:2304.07193*, 2023. 5
- [75] Kunchang Li, Yali Wang, Yizhuo Li, Yi Wang, Yinan He, Limin Wang, and Yu Qiao. Unmasked teacher: Towards training-efficient video foundation models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19948–19960, 2023. 5
- [76] Pritam Sarkar and Ali Etemad. Self-alignment of large video language models with refined regularized preference optimization. *arXiv* preprint arXiv:2504.12083, 2025. 5
- [77] Zhijian Liu, Ligeng Zhu, Baifeng Shi, Zhuoyang Zhang, Yuming Lou, Shang Yang, Haocheng Xi, Shiyi Cao, Yuxian Gu, Dacheng Li, et al. Nvila: Efficient frontier visual language models. *arXiv preprint arXiv:2412.04468*, 2024. 5, 6, 19
- [78] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, et al. Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800, 2024. 6, 19
- [79] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. Gpt-4o system card. *arXiv* preprint arXiv:2410.21276, 2024. 6, 15, 19

Appendix

A Additional Results on VCRBench

We present the detailed results of LVLMs across varying numbers of causal steps in Table S1 and their performance across different sub-categories of VCRBench in Table S2.

Table S1: Performance across videos with varying numbers causal steps. LVLM performance drops sharply as the number of causal steps increases, highlighting challenges in handling complex video-based long-form causal reasoning tasks.

Models]	Number of causal steps						
	3	4	5	6	7			
InternVL2.5 _{38B} [26]	15.9	10.5	10.0	3.3	0.0	11.0		
InternVL2.5 _{78B} [26]	18.2	19.1	8.8	6.7	4.2	14.5		
Qwen2.5-VL-Instruct _{72B} [28]	40.5	37.1	16.2	6.7	4.2	29.0		
GPT4o [79]	33.3	40.0	20.0	13.3	8.3	29.0		
Gemini-1.5-Pro [24]	60.3	50.5	36.2	43.3	20.8	48.2		
Gemini-2.0-Flash-Thinking [24]	64.8	66.7	46.2	53.3	29.2	58.0		

Table S2: Detailed results across the sub-categories of VCRBench.

Models	Add Oil to Your Car	Build Simple Floating Shelves	Change a Tire	Grill Steak	Make Banana Ice Cream	Make French Strawberry Cake	Make French Toast	Make Irish Coffee	Make Jello Shots	Make Lemonade	Make Meringue	Make Pancakes	Overall
InternVL2.5 _{38B} [26]	6.5	0.0	8.0	14.3	16.7	16.1	0.0	20.0	16.7	9.4	7.7	15.6	11.0
InternVL2.5 _{78B} [26]	3.2	6.9	0.0	11.4	33.3	19.4	17.6	10.0	10.0	15.6	11.5	31.2	14.5
Qwen2.5-VL-Instruct _{72B} [28]	16.1	34.5	4.0	22.9	43.3	48.4	20.6	33.3	26.7	40.6	19.2	34.4	29.0
GPT4o [79]	22.6	17.2	4.0	17.1	70.0	32.3	20.6	13.3	43.3	28.1	38.5	40.6	29.0
Gemini-1.5-Pro [24]	12.9	34.5	0.0	45.7	93.3	58.1	47.1	53.3	53.3	62.5	34.6	71.9	48.2
Gemini-2.0-Flash-Thinking [24]	38.7	55.2	12.0	65.7	80.0	61.3	52.9	65.5	66.7	59.4	65.4	65.6	58.0

B Additional Details of VCRBench

B.1 Details of procedures

Table S3 lists the sub-categories included in VCRBench, along with the corresponding number of causal steps and number of samples for each sub-category. The number of causal steps varies across procedures to preserve meaningful causal relationships between intermediate steps and due to the natural diversity in how real-world procedural tasks are performed.

B.2 Details of human-level performance

To obtain human-level performance on VCRBench, we recruit eight volunteers, who are undergraduate or graduate students. We collect their response on a representative subset of roughly 40% of the videos and report the overall performance. An example of user interface is shown in Figure S1.

B.3 Licenses of existing assets used

The videos used in constructing VCRBench are sourced from CrossTask [63] dataset, which is released under BSD-3-Clause license, available here: https://github.com/DmZhukov/CrossTask?tab=BSD-3-Clause-1-ov-file.

Table S3: Varying number of causal steps for different sub-categories of VCRBench.

Sub-categories	# causal steps	# samples
Add Oil to Your Car	3, 4, 5	31
Build Simple Floating Shelves	3, 4, 5	29
Change a Tire	4, 5, 6, 7	25
Grill Steak	3, 4, 5, 6, 7	35
Make Banana Ice Cream	3, 4	30
Make French Strawberry Cake	3, 4	31
Make French Toast	5, 6, 7	34
Make Irish Coffee	3	30
Make Jello Shots	3	30
Make Lemonade	3, 4, 5	32
Make Meringue	4, 5	26
Make Pancakes	3, 4, 5	32

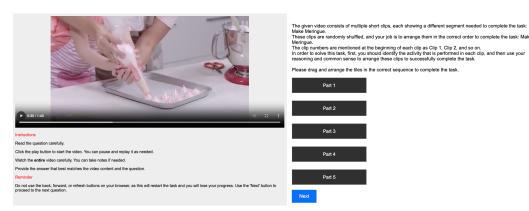


Figure S1: User interface for human evaluation. Questions are randomly shuffled to avoid any potential bias.

B.4 Details of LVLMs

The links to access LVLMs studied in this work are presented in Table S4.

C Additional Details of RRD

The complete instructions used in various RRD setups are mentioned in Figures S2 to S5.

D Broader Impact

Our proposed benchmark advances long-form causal reasoning in LVLMs, which is a critical yet underexplored area. Its novel task formulation avoids common linguistic shortcuts seen in multiple-choice and binary QA formats, allowing for a more reliable assessment of visual understanding. This design can generalize to other vision tasks to enable more accurate evaluation of LVLMs. We hope our benchmark will inspire further research on video-based reasoning and do not anticipate any new negative societal impacts resulting from this work.

```
# Instruction used in Video Recognition step.
The video contains multiple short clips.
The clip numbers are mentioned at the beginning of each clip as Clip 1, Clip 2,
and so on.
Watch each clip carefully, paying attention to its fine-grained actions and
events.
Note the unique events in each clip compared to the rest of the video.
Respond with a one sentence description indicating the key and fine-grained
actions or events for each clip.
Please respond in this format:
Clip 1: <Write one sentence description>
Clip 2: <Write one sentence description>
Your response must not contain anything else.
# Instruction used in Causal Reasoning step.
The following steps are needed to complete the task: {name of the procedure}.
However, these steps are randomly shuffled, and your job is to arrange them in
the correct order to complete the task.
Use your reasoning and common sense to arrange these steps to successfully
complete the task.
{clip descriptions}
The final output should be in this format:
Correct order: <mention the step numbers separated by a comma>
```

Figure S2: Instructions used in video recognition (all-at-once) and causal reasoning (all-at-once) setup.

```
# Instruction used in Video Recognition step.
The video contains multiple short clips.
The clip numbers are mentioned at the beginning of each clip as Clip 1, Clip 2,
and so on.
Watch Clip {step} carefully, paying attention to its fine-grained actions and
Note the unique events in Clip {step} compared to the rest of the video.
Respond with a one sentence description indicating the key and fine-grained
actions or events.
Your response must not contain anything else.
# Instruction used in Causal Reasoning step.
The following steps are needed to complete the task: {name of the procedure}.
However, these steps are randomly shuffled, and your job is to arrange them in
the correct order to complete the task.
Use your reasoning and common sense to arrange these steps to successfully
complete the task.
{clip descriptions}
The final output should be in this format:
Correct order: <mention the step numbers separated by a comma>
```

Figure S3: Instructions used in video recognition (sequential) and causal reasoning (all-at-once) setup.

```
# Instruction used in Video Recognition step.
The video contains multiple short clips.
The clip numbers are mentioned at the beginning of each clip as Clip 1, Clip 2,
Watch each clip carefully, paying attention to its fine-grained actions and
events.
Note the unique events in each clip compared to the rest of the video.
Respond with a one sentence description indicating the key and fine-grained
actions or events for each clip.
Please respond in this format:
Clip 1: <Write one sentence description>
Clip 2: <Write one sentence description>
Your response must not contain anything else.
# Instruction used in Causal Reasoning step.
Here are two intermediate steps to achieving {name of the procedure}:
Event A: {description of one clip}
Event B: {description of another clip}
Which event should occur first?
Pay attention to the causality of events.
Respond with A if Event A should happen first.
Respond with B if Event B should happen first.
Do not provide any other response.
```

Figure S4: Instructions used in video recognition (all-at-once) and causal reasoning (sequential) setup.

```
# Instruction used in Video Recognition step.
The video contains multiple short clips.
The clip numbers are mentioned at the beginning of each clip as Clip 1, Clip 2,
and so on.
Watch Clip {step} carefully, paying attention to its fine-grained actions and
events.
Note the unique events in Clip {step} compared to the rest of the video.
Respond with a one sentence description indicating the key and fine-grained
actions or events.
Your response must not contain anything else.
# Instruction used in Causal Reasoning step.
Here are two intermediate steps to achieving {name of the procedure}:
Event A: {description of one clip}
Event B: {description of another clip}
Which event should occur first?
Pay attention to the causality of events.
Respond with A if Event A should happen first.
Respond with B if Event B should happen first.
Do not provide any other response.
```

Figure S5: Instructions used in video recognition (sequential) and causal reasoning (sequential) setup.

Table S4: Details of LVLMs evaluated on VCRBench.

Models	Weights
InternVL2.5 _{1B} [26]	https://huggingface.co/OpenGVLab/InternVL2_5-1B
InternVL2.5 _{2B} [26]	https://huggingface.co/OpenGVLab/InternVL2_5-2B
LongVU _{3B} [14]	https://huggingface.co/Vision-CAIR/LongVU_Llama3_2_3B
InternVL2.5 _{4B} [26]	https://huggingface.co/OpenGVLab/InternVL2_5-1B
VideoChat2 _{7B} [16]	https://huggingface.co/OpenGVLab/VideoChat2_stage3_Mistral_7B
InternVL2.5 _{8B} [26]	https://huggingface.co/OpenGVLab/InternVL2_5-1B
LLaVA-NeXT-Video7B [22]	https://huggingface.co/LVLMs-lab/LLaVA-Video-7B-Qwen2
MiniCPM-o-V 2.6 _{7B} [78]	https://huggingface.co/openbmb/MiniCPM-o-2_6
Qwen2.5-VL-Instruct _{7B} [28]	https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
VideoLLaMA3 _{7B} [19]	https://huggingface.co/DAMO-NLP-SG/VideoLLaMA3-7B
LongVILA _{7B} [21]	https://huggingface.co/Efficient-Large-Model/qwen2-7b-longvila-256f
LongVU _{7B} [14]	https://huggingface.co/Vision-CAIR/LongVU_Qwen2_7B
NVILA _{15B} [77]	https://huggingface.co/Efficient-Large-Model/NVILA-15B
InternVL2.5 _{26B} [26]	https://huggingface.co/OpenGVLab/InternVL2_5-26B
InternVL2.5 _{38B} [26]	https://huggingface.co/OpenGVLab/InternVL2_5-38B
LLaVA-NeXT-Video _{72B} [22]	https://huggingface.co/LVLMs-lab/LLaVA-Video-72B-Qwen2
Qwen2.5-VL-Instruct _{72B} [28]	https://huggingface.co/Qwen/Qwen2.5-VL-72B-Instruct
InternVL2.5 _{78B} [26]	https://huggingface.co/OpenGVLab/InternVL2_5-78B
GPT4o [79]	Accessed between Jan 2025 to Mar 2025 (gpt-4o-2024-11-20)
Gemini-1.5-Pro [24]	Accessed between Jan 2025 to Mar 2025 (gemini-1.5-pro)
Gemini-2.0-Flash-Thinking [24	Accessed between Jan 2025 to Mar 2025 (gemini-2.0-flash-thinking-exp)