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

Pritam Sarkar
Queen's University, Canada
pritam.sarkar@queensu.ca

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

Website Code Data

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 assess the true visual understanding of LVLMs by preventing them from exploiting linguistic shortcuts, as in multiple-choice or binary QA formats, while also avoiding the challenges associated with evaluating open-ended QA. Our evaluation of state-of-the-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 sub-tasks of video recognition and causal reasoning. Our experiments show that RRD significantly boosts accuracy on VCRBench, with gains of 12.6% *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 even when tackling video-based 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 [34, 55], embodied AI agents [10, 11, 46], spatial intelligence systems [8, 18, 24], and assistive technologies [2, 41], all of which rely on reasoning about causally dependent visual events. While recent advances in vision-language modeling [50, 65, 75] have led to the development of powerful Large Video Language Models (LVLMs) [4, 5, 12–15, 36, 38, 39, 51, 52, 56, 71, 76, 78], 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 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 long-form Causal Reasoning Benchmark (VCRBench)*** 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 [9, 11, 31, 33], VCRBench

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.*

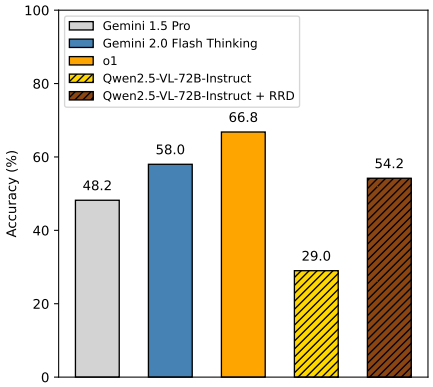

Figure 2. **Impact of RRD.** Qwen2.5-VL-Instruct$_{72B}$ with RRD outperforms Gemini-1.5-Pro and achieve comparable performance to other reasoning specialized models.

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 30%. Further analysis reveals that while these models can often recognize and localize 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 Decomposition (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 performance comparable to that of the current top-performing, reasoning-specialized closed-source models (see Figure 2).

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. Our analysis on various state-of-the-art LVLMs reveals that current LVLMs struggle with long-form causal reasoning due to their inability to 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 [4, 5, 12–15, 36, 38, 39, 51, 52, 56, 71, 76, 78]. 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 [12, 77], dynamic projection techniques that drop redundant frames based on visual similarity [56, 57], and query-based projectors that selectively attend to relevant visual content [36, 37, 59, 65]. In addition to architectural differences, LVLMs vary in their use of vision encoders, ranging from single to multi-encoder setups (e.g., video + image) [44], and from vision-only to vision-language pretrained models [56]. 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 [45, 60]. 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 [74], MSRVTTQA [70], MSVDQA [70], NextQA [69], TGIFQA [27]), comprehensive video understanding (e.g., MVBench [36], TVBench [17], VideoMME [19]), fine-grained temporal understanding (e.g., TVBench [17], TempCompass [42], TemporalBench [6]), long-video understanding (e.g., MLVU [79], LongVideoBench [68]), egocentric video understanding (e.g., Egoschema [47]), and video hallucination (e.g., VideoHallucer [66], HallusionBench [21]), among others. There also exist a few benchmarks focused on video-based reasoning, such as SOK-Bench [63], MMWorld [23], and VILMA [29]. However, a significant gap remains in the evaluation of video-based *causal reasoning* tasks. While some benchmarks address intent (e.g., IntentQA [33]), causal question answering (e.g., Causal-VidQA [31]), or goal-oriented question answering (e.g., EgoPlan-Bench [11], ReXTime [9]), 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 [67]. This paradigm has been further strengthened by post-training alignment techniques such as Reinforcement Learning with Human Feedback (RLHF) [16, 54], which optimize models to generate more helpful and aligned responses. More recent methods like DeepSeek's R1 [22] 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 [64], which aggregate multiple generated responses for robustness, best-of-N sampling [48, 58], which selects the highest-quality sample from multiple candidates, and decomposed prompting [30], 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 preparing pancakes (see Figure 4 for the list of all procedures). For each procedure, we source instructional videos from the CrossTask dataset [80], 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 (depicted in Figure 3) 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[1] as a reference, 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 performed by human annotators to ensure accurate assessment.

• **Step 3.** The selected clips are then randomly shuffled, with the constraint that the original 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

[1] https://www.wikihow.com/

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.

Preparing the questions. A key challenge in evaluating LVLMs is designing a reliable evaluation protocol to correctly assess their true visual capabilities. Most existing video benchmarks [11, 17, 19, 21, 27, 33, 36, 42, 47, 66, 68–70, 74] rely on multiple-choice or binary question-answering formats, to streamline their automated evaluation. However, such setups 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 [1]) as a judge for automated evaluation, a strategy proven to be unreliable and ambiguous in prior work [17]. VCRBench addresses 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 default question template is mentioned in Figure S1.

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 4.

3.2. Evaluation Metrics

We measure the performance of LVLMs on VCRBench with two metrics: overall accuracy and step accuracy [7]. 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:

$$\text{Overall Accuracy} = \frac{\sum_{(q,v)\in\mathcal{D}} \mathbb{1}\left(\pi(q,v)\right)}{|\mathcal{D}|} \text{ and}$$

$$\text{Step Accuracy} = \frac{\sum_{(q,v)\in\mathcal{D}} \frac{\sum_s \mathbb{1}(\pi(q,v))}{s}}{|\mathcal{D}|},$$

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 [20, 61, 62], Mistral [28], and Qwen [3, 72]) with sizes from 1B to 78B parameters for open-source LVLMs; cross-modal adapters (QFormer [32], MLP projector [40], and spatio-temporal compressor [15, 56]); vision encoders with single or dual configurations (CLIP [50], SigLIP [75], DINO [49], and UMT [35]); training methodologies (single-stage or multi-stage) including alignment finetuning for improved reasoning ([5, 53]); and visual frame processing capabilities ranging from 8 (NVILA [43]) to over 500 frames (LongVU [56], Qwen2.5-VL [5]). 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.

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 C.

LVLMs lack video-based long-form causal reasoning.

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 5. 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 [73], 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, o1. We further conduct experiments with different prompt templates to rule out the possibility that the observed

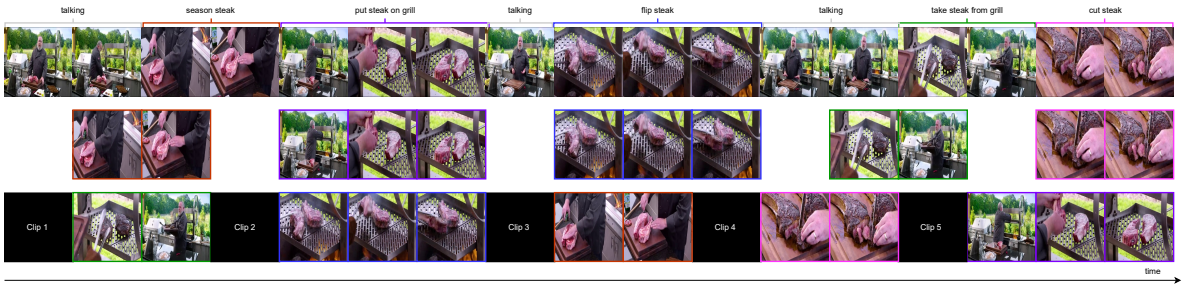

Figure 3. **Overview of video construction. Step 1:** Given a complete video, key procedural steps are identified based on human-annotated timestamps. **Step 2:** We keep the key events and discard those that do not depict visual events directly associated with the goal, such as talking or narrating 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.

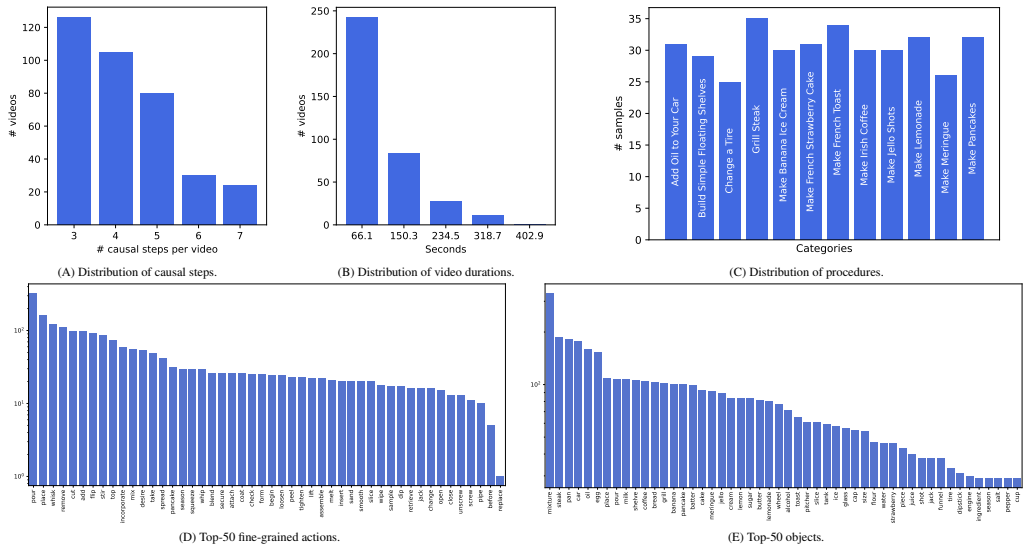

Figure 4. Key **statistics** of our VCRBench.

performance drop arises from instruction misalignment. As shown in Table 2, our default prompt template yields the most stable performance. Interestingly, GPT-4o performs the worst among the closed-source models, achieving a performance similar to Qwen2.5-VL-Instruct$_{72B}$. Overall, even the best-performing model, o1, falls substantially short of human-level performance (66.8% vs. 96.4%). We present sample responses from top-performing LVLMs in Figure 6.

Recognizing and localizing events are not enough: LVLMs lack associative understanding of visual events.

To better understand the limitations of current LVLMs, we conduct additional experiments evaluating their ability to (*i*) recognize and (*ii*) localize intermediate events, both prerequisites for long-form causal reasoning. First, we design an event recognition task where, given a set of intermediate events, LVLMs are asked to identify the correct actions from a mixture of correct and incorrect choices (see Figure 7). Second, we design an event localization task where, given the entire video (as in the long-form causal reasoning setup) and an action name, we evaluate whether LVLMs can correctly locate the corresponding clip for that event (see Figure 8). Results in Table 3 suggest that while LVLMs can recognize and localize individual visual events when provided answer choices, they struggle associating various visual events meaningfully toward a goal, which is required to successfully perform multi-step causal reasoning tasks.

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

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	# Frames	Overall	Step
Random Guess		7.8	24.1
InternVL2.5$_{1B}$ [13]	64	1.4	10.3
InternVL2.5$_{2B}$ [13]	64	6.3	16.2
LongVU$_{3B}$ [56]	1fps	0.0	7.0
InternVL2.5$_{4B}$ [13]	64	1.6	9.5
VideoChat2$_{7B}$ [36]	16	0.3	5.8
InternVL2.5$_{8B}$ [13]	64	2.7	11.1
LLaVA-NeXT-Video$_{7B}$ [78]	64	0.0	17.4
MiniCPM-o-V 2.6$_{7B}$ [73]	64	2.5	11.0
Qwen2.5-VL-Instruct$_{7B}$ [5]	1fps	7.1	20.9
VideoLLaMA3$_{7B}$ [76]	128	1.6	13.1
LongVILA$_{7B}$ [12]	128	0.3	1.1
LongVU$_{7B}$ [56]	1fps	0.0	2.4
NVILA$_{15B}$ [43]	8	0.6	3.6
InternVL2.5$_{26B}$ [13]	64	2.7	13.7
InternVL2.5$_{38B}$ [13]	64	11.0	27.4
LLaVA-NeXT-Video$_{72B}$ [78]	32	5.2	18.6
Qwen2.5-VL-Instruct$_{72B}$ [5]	1fps	29.0	44.0
InternVL2.5$_{78B}$ [13]	64	14.5	34.0
GPT4o [25]	32	29.0	36.6
Gemini-1.5-Pro [51]	1fps	48.2	65.3
Gemini-2.0-Flash-Thinking [51]	1fps	58.0	67.7
o1 [26]	32	66.8	70.2
Human		**96.4**	**98.3**

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

Figure 5. **Failure examples.** Several open-source LVLMs merely list consecutive numbers as the predicted order, exhibiting inability to make a meaningful attempt in VCRBench tasks.

Table 2. **Exploring various prompt templates.** Our default variant (Template 1) achieves an overall stable performance across models.

Models	Template 1	Template 2	Template 3
Qwen2.5-VL-Instruct$_{7B}$	7.1	**8.2**	5.8
Qwen2.5-VL-Instruct$_{72B}$	29.0	19.5	**31.5**

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 sub-tasks are described below:

• **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 (a simplified version; see the exact in Appendix D) :

```
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


```
# Response from o1
A straightforward way to see the correct sequence is:
• First, the boards must be cut to size (Clip 2 shows all the table-saw
and miter-saw work).
• Next, the shelf \box" is assembled and finished (Clip 3 shows gluing,
brad-nailing, and staining).
• Finally, the shelf is mounted on the wall (Clip 1 shows drilling holes,
installing anchors/brackets, and sliding the shelf into place).
So the proper order is:
Correct order:  2, 3, 1
--------------------------------------------------------------------------------
# Response from Gemini-2.0-Flash-Thinking
Rationale:
To build simple floating shelves, the process generally follows these
steps:
1.  Prepare the wood:  This involves cutting the wood to the required
sizes and sanding it to create a smooth surface for finishing.  Clip 5
shows the sanding process of the wooden components.  Although cutting
wood to size isn't shown as a dedicated clip, it is implied and precedes
assembly.  Clip 3 partially shows cutting wood on a table saw and chop
saw.
2.  Assemble the shelf structure:  Once the wood pieces are prepared,
the next step is to assemble the shelf frame.  Clip 3 demonstrates
this step by showing the use of wood glue and a nail gun to join wooden
pieces to form the shelf's inner support structure.
3.  Finish the shelf:  After assembly, the shelf is ready for finishing,
such as staining or painting.  Clip 4 shows the staining process, where
wood stain is applied to the shelves to enhance their appearance.
4.  Install the shelves:  With the shelves built and finished, the
next step is to mount them on the wall.  Clip 1 clearly shows the
installation process, including marking positions, drilling holes,
inserting wall plugs, attaching wooden supports to the wall, and finally
placing the shelf onto these supports.
5.  Showcase the final product:  The final step is to display the
completed shelves, often decorated with items.  Clip 2 shows the
finished floating shelves installed and decorated, with a person
presenting them, indicating this is the final outcome.
Arranging the clips in this logical order creates a step-by-step guide
on how to build simple floating shelves.
Correct order:  Clip 5, Clip 3, Clip 4, Clip 1, Clip 2
--------------------------------------------------------------------------------
# Response from Gemini-1.5-Pro
Certainly!  Here's the correct order of clips to build the simple
floating shelves, based on the logical progression of the project:
Correct order:  2, 3, 4, 5, 1
--------------------------------------------------------------------------------
# Response from Qwen2.5-VL-Instruct72B
Correct order:  Clip 2, Clip 3, Clip 4, Clip 5, Clip 1
```

Figure 6. **Failure examples.** We observe that o1 missed several steps due to its limited capacity to process visual frames. 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 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 and Qwen2.5-VL$_{72B}$ 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.

Question: Which caption matches the video better?
Options:
(A) Squeeze lemon into the pitcher.
(B) Pour lemon juice and water into the pitcher, and add sugar to the lemonade.
(C) Cut lemon into slices.

Figure 7. An event **recognition** task derived from VCRBench.

Question: The given video comprises several clips, identify the clip number that depicts the action: Cut lemon into slices.
Options: (A) Clip 1 (A) Clip 2 (A) Clip 3 (A) Clip 4 (A) Clip 5

Figure 8. An event **localization** task derived from VCRBench.

Table 3. **Limitations of current LVLMs.** LVLMs can recognize and localize individual events fairly well, but struggle to connect them meaningfully toward a specific goal, required in multi-step causal reasoning of VCRBench.

Models	Event Recognition	Event Localization	Long-form Causal Reasoning
Random	33.7	23.5	7.8
LLaVA-NeXT-Video$_{7B}$	84.8	56.7	0.0
InternVL2.5$_{8B}$	64.2	42.3	2.7
Qwen2.5-VL-Instruct$_{7B}$	59.0	58.1	7.1
LLaVA-NeXT-Video$_{72B}$	85.8	60.9	5.2
InternVL2.5$_{78B}$	79.0	70.3	14.5
Qwen2.5-VL-Instruct$_{72B}$	83.1	73.7	29.0

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 (a simplified version; see the exact in Appendix D) :

```
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

Table 4. **Impact of RRD.** Our proposed task decomposition significantly enhances the long-form causal reasoning capabilities of LVLMs, yielding accuracy boosts between 12.6% and 20.9%.

Models	Overall Acc.	Step Acc.
Qwen2.5-VL-Instruct$_{7B}$ [5]	7.1	20.9
+ **RRD (Ours)**	22.5$_{\uparrow 15.4}$	37.3$_{\uparrow 16.4}$
InternVL2.5$_{38B}$ [13]	11.0	27.4
+ **RRD (Ours)**	23.6$_{\uparrow 12.6}$	34.3$_{\uparrow 6.9}$
Qwen2.5-VL-Instruct$_{72B}$ [5]	29.0	43.0
+ **RRD (Ours)**	**49.9**$_{\uparrow 20.9}$	**63.4**$_{\uparrow 20.4}$
InternVL2.5$_{78B}$ [13]	14.5	34.0
+ **RRD (Ours)**	28.2$_{\uparrow 13.7}$	43.5$_{\uparrow 9.5}$

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 4 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 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 5 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. We also evaluate the effect of varying the number of input frames on Qwen2.5-VL-Instruct, which supports longer frame sequences unlike InternVL2.5 (see Table 6). By default, Qwen2.5-VL operates at a sampling rate of 1 fps. Lowering the rate to 0.5 fps leads to a $5-7\%$ performance drop for both the 7B and 72B variants, due to the

Table 5. **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	Overall Acc.	Step Acc.
🎥	📄	49.9	63.4
🎥	🎥 + 📄	46.6$_{\downarrow 3.3}$	62.8$_{\downarrow 0.6}$

Table 6. **Experimenting with varying number of frames.** Both the variants perform better at the default sampling rate of 1fps.

Model	0.5 fps	1.0 fps	2.0 fps
Qwen2.5-VL-Instruct$_{7B}$	14.3	21.9	15.9
Qwen2.5-VL-Instruct$_{72B}$	44.9	49.9	OOM

Table 7. **The effect of further decomposition** of video recognition and causal reasoning steps of RRD. Results are based on Qwen2.5-VL-Instruct$_{72B}$.

Video Recognition	Causal Reasoning	Overall 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	**54.2**	65.1
sequential	sequential	51.0	**66.6**

loss of fine-grained temporal information. However, oversampling at 2 fps does not yield further gains for Qwen2.5-VL-Instruct$_{7B}$. The 72B variant could not be evaluated at 2 fps due to OOM errors, even on a 4×A100 80 GB node.

Sequential recognition improves performance by focusing on one key event at a time.

We conduct a thorough analysis in the main design of RRD. **First**, 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 further simplifies the video recognition task and allows the LVLM to focus on localized analysis of one clip at a time. **Second**, 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 D.

The results are presented in Table 7, 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 relations. 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 8. **Fine-grained results** across different number of causal steps. Qwen2.5-VL-Instruct$_{72B}$ equipped with RRD outperforms Gemini-1.5-Pro and achieve a comparable performance to reasoning specialized closed-source models Gemini-2.0 and o1.

Models	# causal steps (avg. duration in sec.) **3** (85)	**4** (110)	**5** (110)	**6** (110)	**7** (116)	**All**
Qwen2.5-VL-Instruct$_{72B}$	40.5	37.1	16.2	6.7	4.2	29.0
+ **RRD (Ours)**	64.3 (↑ 19.8)	**73.3** (↑ 36.2)	31.2 (↑ 15.0)	23.3 (↑ 16.6)	**33.3** (↑ 29.1)	54.2 (↑ 25.2)
GPT4o	33.3	40.0	20.0	13.3	8.3	29.0
Gemini-1.5-Pro	60.3	50.5	36.2	43.3	20.8	48.2
Gemini-2.0-Flash-Thinking	64.8	66.7	46.2	53.3	29.2	58.0
o1	**79.4**	71.4	**50.0**	**70.0**	**33.3**	**66.8**

RRD effectively bridges the gap between current closed-source and open-source models on VCRBench.

Table 8 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 specialized closed-source models Gemini-2.0-Flash-Thinking and o1. In some setups, it even surpasses or match them.

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. Moreover, our evaluation using VCRBench highlights a sharp performance gap between the best open-source and closed-source models. As an initial step towards enabling such capabilities in open-source models, we introduce RRD, a simple approach that decomposes video-based causal reasoning into video recognition and causal reasoning tasks. RRD significantly improves multi-step causal reasoning of LVLMs, for instance, Qwen2.5-VL-Instruct$_{72B}$ with RRD outperforms Gemini-1.5-Pro and achieves performance comparable to that of reasoning specialized closed-source models.

Acknowledgment

We thank Debaditya Shome and Nishq Poorav Desai for their help in building the platform used to collect human-

level performance. We also thank members of our lab who participated in obtaining human-level performance in VCR-Bench. We also thank the Bank of Montreal and Mitacs for funding this research, and the Vector Institute for providing computational resources.

References

[1] 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

[2] 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

[3] 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. 4

[4] 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

[5] 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, 4, 6, 7

[6] 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

[7] 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. 4

[8] 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

[9] 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. 1, 3

[10] 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

[11] 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, 3, 4

[12] 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, 4

[13] 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. 6, 7, 4

[14] 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.

[15] 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, 4

[16] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *NeurIPS*, 30, 2017. 3

[17] Daniel Cores, Michael Dorkenwald, Manuel Mucientes, Cees GM Snoek, and Yuki M Asano. Tvbench: Redesigning video-language evaluation. *arXiv preprint arXiv:2410.07752*, 2024. 3, 4

[18] 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

[19] 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

[20] 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. 4

[21] 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

[22] 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

[23] 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

[24] 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

[25] 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, 4

[26] Aaron Jaech, Adam Kalai, Adam Lerer, Adam Richardson, Ahmed El-Kishky, Aiden Low, Alec Helyar, Aleksander Madry, Alex Beutel, Alex Carney, et al. Openai o1 system card. *arXiv preprint arXiv:2412.16720*, 2024. 6, 4

[27] 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

[28] Fengqing Jiang. Identifying and mitigating vulnerabilities in llm-integrated applications. Master's thesis, University of Washington, 2024. 4

[29] 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

[30] 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

[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. 1, 3

[32] 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. 4

[33] 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. 1, 3, 4

[34] 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

[35] 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. 4

[36] 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

[37] 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

[38] 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

[39] 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

[40] 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. 4

[41] 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

[42] 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

[43] 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. 4, 6

[44] 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

[45] 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. 3

[46] 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

[47] 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

[48] 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

[49] 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. 4

[50] 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, 4

[51] 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, 4

[52] 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

[53] Pritam Sarkar and Ali Etemad. Self-alignment of large video language models with refined regularized preference optimization. *arXiv preprint arXiv:2504.12083*, 2025. 4

[54] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017. 3

[55] 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

[56] 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, 4, 6

[57] 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

[58] 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. 3

[59] 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

[60] 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

[61] 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. 4

[62] 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. 4

[63] 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

[64] 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

[65] 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

[66] 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

[67] 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

[68] 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

[69] 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

[70] 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

[71] 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

[72] 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. 4

[73] 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. 4, 6

[74] 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*, pages 9127–9134, 2019. 3, 4

[75] 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, 4

[76] 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, 4

[77] 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

[78] 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, 4

[79] 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

[80] 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. 3, 2

Supplementary Material

A. Release Statement

In compliance with the CVPR 2026 author guidelines, we are not permitted to include any expandable private or anonymous links, even those hosted on *anonymous.4open.science* or private *Kaggle* repositories, for sharing data, code, or model weights during the review phase. As a result, we could not provide direct links to the full dataset or source code. Furthermore, since VCRBench contains videos exceeding the 200 MB supplementary file limit, the entire benchmark cannot be added as a supplementary file. Under these restrictions, we include a subset of VCRBench and the associated code in the supplementary materials. Post review phase, we will add the open-source link in the paper to access the entire benchmark and accompanying codebase for transparency, reproducibility, and community accessibility.

B. Limitations and Future Work

While RRD effectively enhances the video-based causal reasoning capabilities of LVLMs, the reasoning primarily occurs in the language domain rather than directly in the visual domain. Future work could focus on developing LVLMs that perform complex reasoning directly from visual inputs, thereby improving both their effectiveness and efficiency. In evaluating performance on VCRBench, we use two metrics: overall accuracy and step accuracy. While overall accuracy effectively captures complete success and failure cases, it does not reflect partial correctness. Our step accuracy metric provides some indication of partial success by measuring the longest contiguous matching subsequence between predicted and ground-truth event sequences. However, it overlooks scenarios where the model produces correct events with intermittent mistakes, or where certain events may carry more importance than others. Designing metrics that account for these nuanced cases remains an open question. We believe that VCRBench offers a strong foundation for further research in this direction.

C. Additional Details of VCRBench

C.1. Instruction Templates

The instruction templates explored in constructing VCRBench are presented in Figures S1 - S3.

```
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.
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 S1. Template 1 (default).

```
You are given a video composed of multiple short clips, each
representing a different step required to complete the task: {name
of the procedure}.
These clips are randomly shuffled. Your task is to determine the
correct chronological order to complete the task successfully.
Each clip is labeled at the beginning (e.g., Clip 1, Clip 2, etc.).
To solve this task, follow these steps:
1. Carefully examine each clip and describe the main activity or event
it shows.
2. Based on your understanding of how the task goal is typically
performed, reason through how these steps would logically follow one
another.
3. Use common sense and temporal reasoning to identify the correct
sequence.
Finally, provide your answer in this format:
Correct order: <mention the Clip numbers separated by a comma>
```

Figure S2. Template 2.

```
You are shown a set of shuffled video clips, each depicting a distinct
step involved in completing a larger task: {name of the procedure}.
Each clip is labeled at the beginning (e.g., Clip 1, Clip 2, etc.).
Your objective is to determine the correct chronological order of the
clips to accurately reconstruct the full task.
To approach this task, proceed as follows:
- First, describe what is happening in each clip and infer what role it
plays in the overall task.
- Then, reason about how these steps are likely connected | which actions
must logically come before or after others.
- Consider real-world knowledge and causality to guide your ordering.
Only after this reasoning, write your final answer in the following
format:
Correct order: <mention the Clip numbers separated by commas>
```

Figure S3. Template 3.

C.2. Human Evaluation

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. Note that the questions are randomly shuffled to avoid any potential bias. The user interface to obtain human performance is shown in Figure S4.

C.3. Licenses of Existing Assets Used

The videos used in constructing VCRBench are sourced from CrossTask [80] dataset, which is released under BSD-3-Clause license, available here: https://github.com/DmZhukov/CrossTask?tab=BSD-3-Clause-1-ov-file.

C.4. Details of LVLMs

The URLs to access LVLMs studied in this work are presented in Table S1.

C.5. Additional Results on VCRBench

We present the detailed results of LVLMs across varying numbers of causal steps in Table S2 and their performance across different sub-categories of VCRBench in Table S3.

D. Additional Details of RRD

The complete instructions used in various RRD setups are mentioned in Figures S5 to S8.

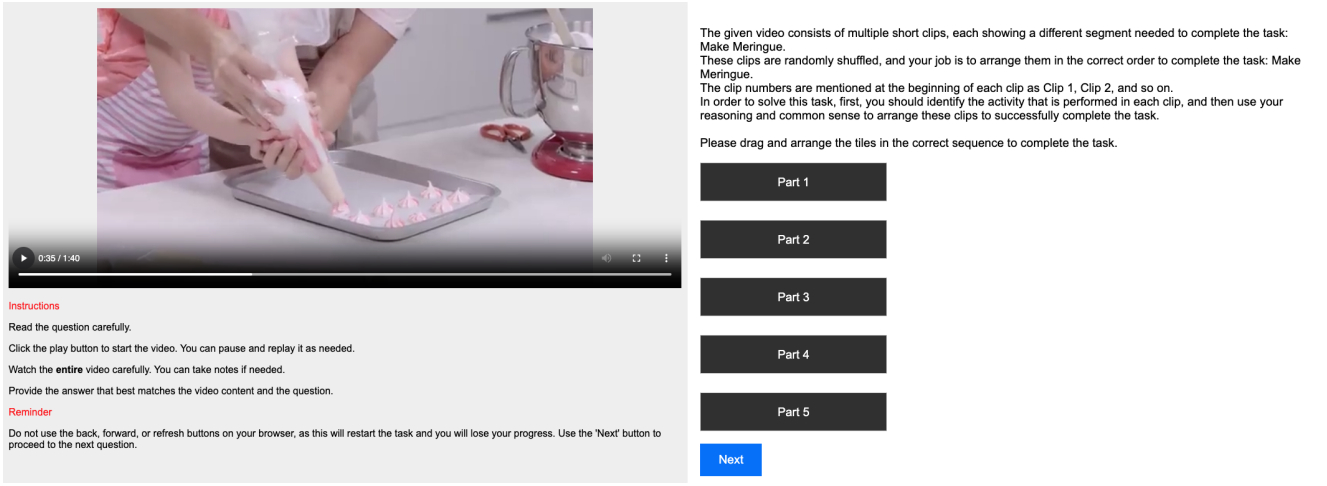

Figure S4. Setup used in collecting human-level performance on VCRBench.

```
# 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 S5. Instructions used in video recognition (all-at-once) and causal reasoning (all-at-once) setup of RRD.

```
# 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.
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 S6. Instructions used in video recognition (sequential) and causal reasoning (all-at-once) setup of RRD.

```
# 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.
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 S7. Instructions used in video recognition (all-at-once) and causal reasoning (sequential) setup of RRD.

```
# 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 S8. Instructions used in video recognition (sequential) and causal reasoning (sequential) setup of RRD.

Table S1. Details of LVLMs evaluated on VCRBench.

Models	Weights
InternVL2.5$_{1B}$ [13]	https://huggingface.co/OpenGVLab/InternVL2_5-1B
InternVL2.5$_{2B}$ [13]	https://huggingface.co/OpenGVLab/InternVL2_5-2B
LongVU$_{3B}$ [56]	https://huggingface.co/Vision-CAIR/LongVU_Llama3_2_3B
InternVL2.5$_{4B}$ [13]	https://huggingface.co/OpenGVLab/InternVL2_5-1B
VideoChat2$_{7B}$ [36]	https://huggingface.co/OpenGVLab/VideoChat2_stage3_Mistral_7B
InternVL2.5$_{8B}$ [13]	https://huggingface.co/OpenGVLab/InternVL2_5-1B
LLaVA-NeXT-Video$_{7B}$ [78]	https://huggingface.co/LVLMs-lab/LLaVA-Video-7B-Qwen2
MiniCPM-o-V 2.6$_{7B}$ [73]	https://huggingface.co/openbmb/MiniCPM-o-2_6
Qwen2.5-VL-Instruct$_{7B}$ [5]	https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct
VideoLLaMA3$_{7B}$ [76]	https://huggingface.co/DAMO-NLP-SG/VideoLLaMA3-7B
LongVILA$_{7B}$ [12]	https://huggingface.co/Efficient-Large-Model/qwen2-7b-longvila-256f
LongVU$_{7B}$ [56]	https://huggingface.co/Vision-CAIR/LongVU_Qwen2_7B
NVILA$_{15B}$ [43]	https://huggingface.co/Efficient-Large-Model/NVILA-15B
InternVL2.5$_{26B}$ [13]	https://huggingface.co/OpenGVLab/InternVL2_5-26B
InternVL2.5$_{38B}$ [13]	https://huggingface.co/OpenGVLab/InternVL2_5-38B
LLaVA-NeXT-Video$_{72B}$ [78]	https://huggingface.co/LVLMs-lab/LLaVA-Video-72B-Qwen2
Qwen2.5-VL-Instruct$_{72B}$ [5]	https://huggingface.co/Qwen/Qwen2.5-VL-72B-Instruct
InternVL2.5$_{78B}$ [13]	https://huggingface.co/OpenGVLab/InternVL2_5-78B
GPT4o [25]	gpt-4o-2024-11-20
Gemini-1.5-Pro [51]	gemini-1.5-pro
Gemini-2.0-Flash-Thinking [51]	gemini-2.0-flash-thinking-exp
o1 [26]	o1-2024-12-17

Table S2. 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					Overall
	3	4	5	6	7	
InternVL2.5$_{38B}$ [13]	15.9	10.5	10.0	3.3	0.0	11.0
InternVL2.5$_{78B}$ [13]	18.2	19.1	8.8	6.7	4.2	14.5
Qwen2.5-VL-Instruct$_{72B}$ [5]	40.5	37.1	16.2	6.7	4.2	29.0
GPT4o [25]	33.3	40.0	20.0	13.3	8.3	29.0
Gemini-1.5-Pro [51]	60.3	50.5	36.2	43.3	20.8	48.2
Gemini-2.0-Flash-Thinking [51]	64.8	66.7	46.2	53.3	29.2	58.0
o1 [26]	79.4	71.4	50.0	70.0	33.3	66.8

Table S3. 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$_{38B}$ [13]	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$_{78B}$ [13]	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$_{72B}$ [5]	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 [25]	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 [51]	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 [51]	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
o1 [26]	35.5	17.2	28.0	68.6	93.3	71.0	70.6	83.3	80.0	84.4	73.1	87.5	66.8