Two Causally Related Needles in a Video Haystack

Miaoyu Li[*], **Qin Chao**[*], **and Boyang Li**
College of Computing and Data Science
Nanyang Technological University, Singapore
{miaoyu.li, chao0009, boyang.li}@ntu.edu.sg

Abstract

Properly evaluating the ability of Video-Language Models (VLMs) to understand long videos remains a challenge. We propose a long-context video understanding benchmark, CAUSAL2NEEDLES, that assesses two crucial abilities insufficiently addressed by existing benchmarks: (1) extracting information from two separate locations (two needles) in a long video and understanding them jointly, and (2) modeling the world in terms of cause and effect in human behaviors. CAUSAL2NEEDLES evaluates these abilities using noncausal one-needle, causal one-needle, and causal two-needle questions. The most complex question type, causal two-needle questions, require extracting information from both the cause and effect events from a long video and the associated narration text. To prevent textual bias, we introduce two complementary question formats: locating the video clip containing the answer, and verbal description of a visual detail from that video clip. Our experiments reveal that models excelling on existing benchmarks struggle with causal 2-needle questions, and the model performance is negatively correlated with the distance between the two needles. These findings highlight critical limitations in current VLMs. The dataset is available at: https://huggingface.co/datasets/causal2needles/Causal2Needles

1 Introduction

On many popular benchmarks [Hendrycks et al., 2020, Cobbe et al., 2021, Srivastava et al., 2022, Chen et al., 2021], recent AI systems achieved performance comparable to humans. However, real-world observations suggest that there is still a significant gap between AI and human capabilities [West et al., 2023, Frieder et al., 2023, Amirizaniani et al., 2024, Chinchure et al., 2025]. This apparent contradiction suggests that (1) there is much overfitting to popular benchmarks such as GSM8K [Mirzadeh et al., 2024] through mechanisms like data contamination [Singh et al., 2024], and (2) existing benchmarks may not fully reflect differences between machine intelligence and human intelligence [Dziri et al., 2023, Wu et al., 2024]. As a result, the development of sophisticated benchmarks that critically evaluate model capabilities has become a high priority for AI research.

Within the context of long video understanding, we investigate two crucial limitations of existing benchmarks for Video-Language Models (VLMs). First, benchmarks evaluating information extraction from a single location do not fully reflect the long video understanding ability of VLMs. A popular type of evaluation for long-context models adopts the "needle in a haystack" problem formulation [Kamradt, 2023, Wang et al., 2024b], where the "needle" represents the information to be extracted from the long context. Some long video benchmarks feature 1-needle questions that require information from a single location in the video. However, research in NLP indicates that model performances on 1-needle questions are often much higher than performances on questions requiring the extraction and understanding multiple needles, revealing the limitations of 1-needle

[*]Equal contribution

39th Conference on Neural Information Processing Systems (NeurIPS 2025) Track on Datasets and Benchmarks.

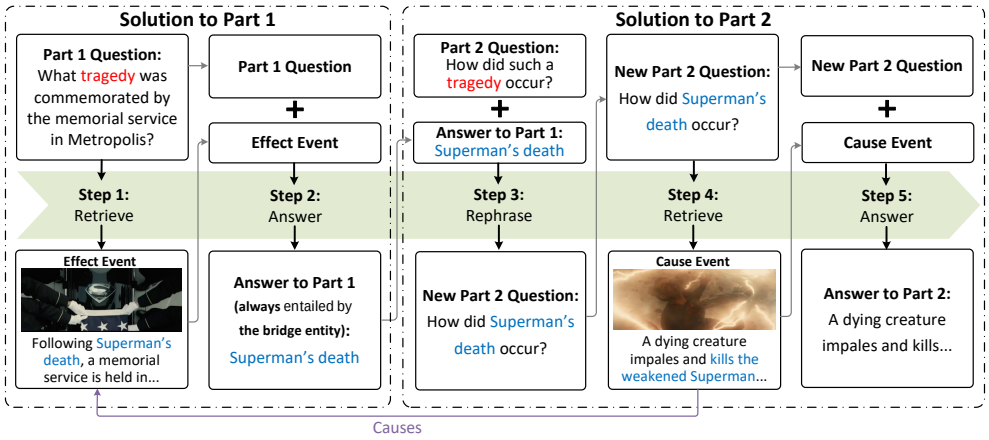

Figure 1: The logical solution process for the 2-needle questions of CAUSAL2NEEDLES. Each step involves an operation, with the input shown above the step and the output below the step. The question purposely refer to the bridge entity, "Superman's death," ambiguously as "tragedy." As a result, one must first resolve the bridge entity using Part 1 before answering Part 2. This question design mandates joint understanding of both the cause and effect events. Note that the steps are necessary only in an information-processing sense. A VLM may adopt different steps.

questions for assessing long-context understanding [Li et al., 2023, Vodrahalli et al., 2024, Yang, 2024, Levy et al., 2024]. However, multiple-needle questions remain rare in the multimodal setting.

Second, existing benchmarks offer incomplete evaluation of whether VLMs possess an internal world model, which captures the underlying mechanisms governing object dynamics and human behaviors and enables predictions of future events from an intervention [Forrester, 1971, Ha and Schmidhuber, 2018]. Existing evaluations of world models [Guan et al., 2024a, Motamed et al., 2025, Kang et al., 2024] focus solely on object motion prediction, neglecting human behaviors and event causality [Sun et al., 2024].

To address these limitations, we propose a long-context video understanding benchmark, CAUSAL2NEEDLES, comprising 2,606 1-needle questions and 1,494 2-needle questions. Out of these, 902 1-needle questions and all 2-needle questions involve causal, world-model reasoning. CAUSAL2NEEDLES evaluates two key abilities of VLMs: (1) extracting relevant information from two locations in long videos and jointly reasoning about them, and (2) modeling the world in terms of causes and effects of human behaviors.

The causal 2-needle questions are constructed from a pair of cause and effect events. Each question should require the VLM to first retrieve the effect event and then the cause event. To formulate the question, we identify a *bridge entity*, which is an entity or a piece of information shared by the cause event and the effect event. Part one of the question asks the VLM to resolve the bridge entity by retrieving the effect event. Part two of the question requires the retrieval of the cause of the effect event. As an example, Fig. 1 shows a cause event and an effect event that share a bridge entity, "Superman's death," ambiguously referred to as "tragedy." To answer question part one, the model must resolve the content of the tragedy by retrieving the effect event, which reveals the tragedy is Superman's death. Only after that, the model can answer question part two by retrieving the video clip showing the cause of Superman's death. If the bridge entity were explicitly stated, the question would become "how did Superman die, leading to his memorial service in Metropolis", which can be answered by retrieving the cause event directly. In that case, the 2-needle question would degenerate to a 1-needle question. This novel problem formulation allows CAUSAL2NEEDLES to assess the two-needle ability and the causal reasoning ability together.

CAUSAL2NEEDLES is constructed on top of movie summary videos, containing video clips from the movies and narration text. However, this may create a shortcut that VLMs can exploit. The video clips could be difficult to understand without the narration text, but if we feed the narration text as input, the VLM may answer the question from the text directly without using the video. This is the infamous phenomenon of textual bias [Ko et al., 2023, Cores et al., 2024, Xiao et al., 2024].

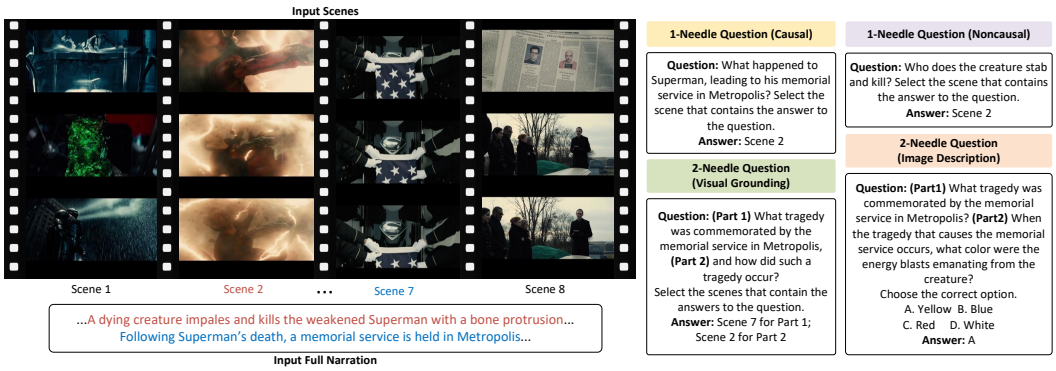

Figure 2: The evaluation framework of CAUSAL2NEEDLES. To help models understand the storyline, we also feed the full textual narration into the model. Four types of questions are designed for each pair of causally related events.

As mitigation, we introduce two complementary question formats. The first format, called *visual grounding*, requires the model to select the video clip corresponding to the event it needs to retrieve. The second format, called *image description*, requires the model to answer multiple-choice questions about the appearance of the retrieved video clip. Fig. 2 shows some examples. Visual grounding questions necessitate the understanding of video clips, but they may underestimate model performance as the format may be out-of-distribution (OOD) to most VLMs. In contrast, image description questions circumvent the OOD issue, but the VLM may benefit from knowledge of the movie learned from pretraining data, such as the color of Lois Lane's outfit, leading to overestimation of model abilities. CAUSAL2NEEDLES adopts the visual grounding format for all 1-needle questions and splits the 2-needle questions between the two formats (747 each).

Experiments reveal several important findings. First, causal questions appear substantially more difficult than noncausal questions. Second, models that perform well on 1-needle questions exhibit remarkable performance drops on 2-needle questions. Finally, the distance between the two needles is negatively correlated with model performance. Taken together, these results demonstrate that joint understanding of and causal reasoning over two separate video locations remain important weaknesses of existing VLMs. We summarize the contributions of this paper as follows:

• We propose the CAUSAL2NEEDLES benchmark that contains causal 1-needle and 2-needle questions, generated from story videos. We devise a question creation strategy utilizing bridge entities to force the video-language models to jointly understand two video clips arbitrarily located in the context window.

• We reveal significant weaknesses in current VLMs in the joint understanding and retrieval of two separate video clips, even though their 1-needle performances are high. The more distant the two clips are, the worse the models perform.

2 Related Work

2.1 Long Video Understanding Benchmarks

Retrieval of information from a specific location in a long video (usually called "needle in a haystack") is widely regarded as a key capability in long-video understanding [Kamradt, 2023, Wang et al., 2024b]. In Tab. 1, we compare recent long-video benchmarks with multiple-needles-in-haystack tasks side by side to highlight the unique contribution of the proposed CAUSAL2NEEDLES dataset. EgoSchema [Mangalam et al., 2023] does not measure specific model capabilities (*e.g.*, 1-needle vs. 2-needle, reasoning vs. recognition) as the questions are not categorized by the required capability, resulting in limited diagnostic precision. MLVU [Zhou et al., 2024] require models to retrieve visually distinct images that are artificially inserted into videos, which may allow for shortcut features due to the domain gap between the artificial needle and the original content. Multi-needle questions also exist in MVBench [Li et al., 2024c] and TVBench [Cores et al., 2024], which contain questions to evaluate the model ability to track an object across multiple temporal locations. Similar questions are also featured in VideoMME [Fu et al., 2024] (*e.g.*, counting problems) and LongVideoBench

Table 1: A comparison of CAUSAL2NEEDLES with other multi-needle long video benchmarks. CAUSAL2NEEDLES is the only benchmark dedicated to needle-in-haystack problems (Diagnostic Precision) that requires joint understanding of the two needles and the identification of cause events from effect events.

Benchmark	Video Length	# QA	Diagnostic Precision	Needle Type	Joint Understanding	Causal Reasoning
EgoScheme [Mangalam et al., 2023]	180 s	5,000	✗	Natural	✗	✓
MVBench [Li et al., 2024c]	16 s-40 s	4,000	✓	Natural	✗	✓
TVBench [Cores et al., 2024]	16 s-40 s	2,525	✓	Natural	✗	✗
MLVU [Zhou et al., 2024]	180 s-3600 s	3,102	✓	Artificial	✗	✗
VideoMME [Fu et al., 2024]	1018s	2,700	✓	Natural	✗	✓
LVB Wang et al. [2024a]	473 s	6,679	✓	Natural	✗	✗
CAUSAL2NEEDLES (Ours)	438 s	4,100	✓	Natural	✓	✓

(*e.g.*, L2-SSS questions) [Wu et al., 2025]. However, questions in the above datasets typically require only independent understanding of each needle. Moreover, the multiple needles often only related by surface-level visual cues, such as matching visual appearances of objects, rather than semantic relations like causality.

In contrast, the CAUSAL2NEEDLES benchmark requires models to jointly understand both the cause and the effect—the understanding of the cause relies on the correct interpretation of the clue in the effect. Furthermore, the two needles are related by causality rather than visual similarity, necessitating semantic understanding. These designs establish a challenging and unique benchmark for long-video understanding.

2.2 World Model

There is an ongoing debate [Gurnee and Tegmark, 2023, Liu et al., 2025, Guan et al., 2024b, Hu et al., 2024, Kang et al., 2024, Motamed et al., 2025] over whether deep neural networks learn to develop internal world models—whether they can identify the underlying principles, such as laws of physics, that govern the observational data they learn from. Such world models, if present, can be used to predict future outcomes of possible interventions, such as kicking a ball or blowing a candle [Forrester, 1971, Ha and Schmidhuber, 2018, Li et al., 2025].

However, this debate largely overlooks causality between events involving humans, which could be induced by factors like biology (e.g., heavy sweating causes thirst), psychology (e.g., hearing compliments makes people feel good), and social norms (e.g., providing good service leads to tipping). Although causal questions sporadically appear in video understanding datasets, such as the episodic reasoning questions in MVBench, it is usually difficult to isolate them and quantitatively measure this factor independently. In contrast, CAUSAL2NEEDLES is dedicated to evaluating reasoning over cause and effect in the context of human behaviors, which allows precise diagnosis of this capability and fills a gap in benchmarks of world models.

3 CAUSAL2NEEDLES: Dataset Construction

CAUSAL2NEEDLES is built on two video-language datasets, YMS [Dogan et al., 2018] and SyMoN [Sun et al., 2022]. With a total of 192 fully annotated movie recap videos, the datasets offer a rich collection of human-behavior events from diverse movie genres. Each event consists of a narration sentence and its corresponding video clip. Details such as the distribution of movie themes and the temporal distance between causal events can be found in Appendix Sec. A.

To facilitate the evaluation of models of diverse capabilities, CAUSAL2NEEDLES consists of both 1-needle and 2-needle questions, as well as causal and noncausal questions, which have different difficulty levels. The generation of both 1-needle and 2-needle questions depend on pairs of cause and effect sentences, extracted in a causal relationship extraction step (Sec. 3.1). We describe the generation of 1-needle questions in Sec. 3.2 and the generation of 2-needle questions in Sec. 3.3.

3.1 Causal Relationships Extraction

We employ a Large Language Model (LLM) to extract causal relationships from narration by combining global and local event graphs, based on an event graph extraction method [Sun et al., 2024]. Each event (a narration sentence) is represented as a node, and causal relationships as directed edges. The global graph is extracted from the complete narration of a video. The graph captures long-range causal relationships but may be incomplete, because LLM often overlook sentences in the middle section of a long context [Liu et al., 2024]. To address this, we introduce a sliding-window approach that extracts local graphs from 15-sentence segments with a 5-sentence stride. These local graphs capture more detailed causal relationships with a shorter range. We merge the global and local graphs to obtain comprehensive and long-range causal relationships. To avoid superficial causal relationships resulting from temporal adjacency, we retain only those where the cause and effect are separated by at least three events. In CAUSAL2NEEDLES, the distance between cause and effect events ranges from 3 to 21 events. Appendix Sec. A and B contain more details. The prompt used is shown in Appendix Fig. 9.

3.2 Generation of One-Needle Questions

With the extracted causal relationships, we use an LLM to generate simple 1-needle questions. To facilitate reader understanding, we illustrate the process using the cause and effect events shown in Fig. 2. We first prompt the LLM to generate noncausal one-needle questions, which ask for a detail in either the cause or the effect event. For the cause event in Fig. 2, the generated question is: "Who does the creature stab and kill?" We can similarly generate another question for the effect event. The causal one-needle questions differ from the noncausal question by requiring understanding of the causal relation between the two events. We prompt the LLM to generate a question asking for the cause event of a specific effect. The generated question for the example in Fig. 2 is: "What happened to Superman, leading to his memorial service in Metropolis?" Finally, to mitigate potential textual bias from the input narration, we combine these questions with the visual grounding instruction: "Select the scene that contains the answer to the question." The prompt for 1-needle question generation is shown in Appendix Fig. 10.

3.3 Generation of Two-Needle Questions

Generating 2-needle questions involves two steps after obtaining causal relationships: (1) rephrasing the cause and effect sentences to establish the bridge entity between them, and (2) generating a question that requires joint understanding of both cause and effect events.

Rephrasing the Cause and Effect Sentences. The generation of 2-needle questions is based on a bridge entity that connects cause and effect events. The bridge entity serves two critical purposes: it provides a hint that drives the model to locate the cause event, so we do not provide the cause event directly; it forces the model to resolve the content of the vaguely phrased bridge entity by locating the effect event, before it can locate the cause event.

However, the bridge entity may not be explicitly mentioned in the original cause and effect sentences. Therefore, we prompt a VLM to rephrase the cause and effect sentences to explicitly establish the bridge entity. For example, the cause sentence, "As it dies, the creature stabs and kills the weakened Superman with one of its bone protrusions," and the effect sentence, "A memorial is held for Superman in Metropolis," are rephrased into sentences in Fig. 2. To maintain semantic consistency between the rephrased sentences and their corresponding clips, we also input the video clips into the VLM as a constraint.

Generation of Visual Grounding Questions. After obtaining causal sentence pairs with a clear bridge entity, we utilize an LLM to generate two parts of a visual grounding 2-needle question. We use the causal relationship in Fig. 2 to demonstrate this process. First, we prompt the LLM to generate a Part 1 question that requires retrieving the effect event to identify the bridge entity. Specifically, we ask it to extract the bridge entity that establishes the causal relationship, such as "Superman's death." Next, we instruct the LLM to rephrase this bridge entity as a vague reference, such as "tragedy." After that, using this vague reference, we prompt the LLM to create a question that uses the effect sentence as context and the bridge entity as the answer. In this case, the final output is: "What tragedy was commemorated by the memorial service in Metropolis?" Further, we

prompt the LLM to generate a Part 2 question, which requires retrieving the cause event. In the example, with the vague reference "tragedy", we instruct the LLM to generate a question whose answer appears exclusively in the cause sentence. The output is: "How did such a tragedy occur?" It is important to keep the bridge entity vague in this question, so that the question does not give the effect event away. Instead, the model must retrieve the video clip of the effect event and resolve the bridge entity reference. Finally, combining the two question parts with a task instruction, we complete a visual grounding 2-needle question: "What tragedy was commemorated by the memorial service in Metropolis, and how did such a tragedy occur? Select the scenes that contain the answers to the question." To answer it, the model must jointly understand cause and effect events and ground the answers on video clips. The prompts are in Appendix Fig. 12.

Generation of Image Description Questions. The visual grounding question format may be out-of-distribution for some models. As a result, it may underestimate their performance. To fix this, we also generate questions in a complementary format: multiple-choice image description questions. These questions share the same Part 1 as visual grounding questions, but modify Part 2 to ask about visual details of the cause event.

With the cause video clip, the cause sentence, and the effect sentence as input, we prompt an LLM to generate a question that asks for an attribute of a visible object in the clip. This question follows the template: "When the event that causes ⟨`Effect Event`⟩ occurs, ⟨`Image Description Question`⟩." For example, in the clip where Metropolis mourns Superman, the LLM output could be: "When the battle that causes Superman's death and leads to the memorial service occurs, what color were the energy blasts emanating from Doomsday?" We then further prompt the LLM to generate four challenging answer options based on the question.

However, the generated questions may reveal the bridge entity or include excessive object details, allowing the model to locate the cause clip without truly understanding the Part 1 question. For instance, the initially generated question reveals the bridge entity "Superman's death" and the name "Doomsday", which reduces the difficulty of locating the video clip. To mitigate this issue, we prompt the LLM to obscure the bridge entity and the named entity being inquired about. We then obtain the final image description 2-needle question, which begins with the Part 1 question: "What tragedy was commemorated by the memorial service in Metropolis? When the tragedy that causes the memorial service occurs, what color were the energy blasts emanating from the creature? A. Yellow B. Blue C. Red D. White" The prompts are shown in Appendix Fig. 14 and Fig. 15.

3.4 Quality Evaluation of Generated Questions

We conduct automatic evaluation and human evaluation of the quality of questions and bridge entities, which provide vital information for 2-needle questions. We evaluate 4 quality factors. Factor 1 is if the bridge entity is truly shared by the cause and effect events it is supposed to connect. We report the proportion of affirmative answers as the final score. Factor 2 is the correctness of purposely vague references to the bridge entities, or if the vague reference is indeed more ambiguous than the original bridge entity but still preserves its core meaning. The evaluation result for each vague reference should be "Yes" or "No". We report the proportion of "Yes". Factor 3 is the factual correctness of questions, or the extent to which a question is consistent with the story and does not introduce hallucination or contradiction to the story. We use a 5-point scale, where 1 is the lowest score and 5 the highest. Factor 4 is the readability of questions, as reflected by the naturalness, grammar, and clarity, on a 5-point scale.

To verify that the evaluation LLMs are not biased to always indicate high quality, we also create several random baselines. For Factor 1, we randomly match the bridge entities and the cause-effect event pairs, which should cause the LLM to answer No to shared existence of the bridge entity. For Factor 2, we randomly shuffle the correspondence between the bridge entities and the vague references. For Factor 3, we randomly shuffle the correspondence between the questions and the cause-effect event pairs. For Factor 4, due to the difficulty in writing unreadable questions, we do not construct any random baseline for readability.

We utilize two state-of-the-art models as the evaluation LLMs, ChatGPT-4.1 and Gemini-2.0-flash (neither is involved in the question generation process), and recruit five human annotators. The models evaluate both 1-needle and 2-needle questions, while the annotators evaluate 136 visual grounding

Table 2: Evaluation results of generated questions. VG and ID refer to visual grounding and image description, respectively. 1-N and 2-N denote 1-needle and 2-needle questions. Numbers in parentheses indicate the performance of random baselines.

Models	Shared Existence of Bridge Entities	Correctness of Vague References	Factual Correctness of Questions				Readability of Questions			
			Noncausal 1-N	Causal 1-N	VG 2-N	ID 2-N	Noncausal 1-N	Causal 1-N	VG 2-N	ID 2-N
ChatGPT-4.1	95.6% (0.3%)	91.0% (3.6%)	4.71 (1.10)	4.62 (1.12)	4.99 (1.10)	4.74 (1.13)	4.91	4.85	4.83	4.67
Gemini-2.0-flash	95.0% (3.8%)	98.7% (2.5%)	4.75 (1.05)	4.66 (1.02)	4.96 (1.01)	4.83 (1.00)	4.75	4.18	4.69	4.25
Human	82.4%	98.5%	-	-	4.50	-	-	-	4.80	-

Table 3: Quantitative results (accuracy, %) of VLMs on our benchmark. "Forward" refers to inputting video clips in chronological order, while "Reverse" uses reverse order. "Avg" denotes results averaged over both orders. Best scores are in **bold**.

Models	Noncausal 1-N Questions	Causal 1-N Questions	VG 2-N Questions									ID 2-N Questions
			Forward			Reverse			Avg			
			Part 1	Part 2	Both	Part 1	Part 2	Both	Part 1	Part 2	Both	
Human	–	78.2	83.7	85.9	79.3		-			-		88.2
Proprietary Models												
ChatGPT-4o	**56.8**	**39.2**	16.7	39.2	9.4	**45.4**	21.2	**13.4**	**31.1**	30.2	**11.4**	59.2
Gemini-1.5-pro	55.4	35.6	**21.0**	**40.0**	**10.2**	35.7	**21.4**	8.4	28.4	**30.7**	9.3	**60.9**
ChatGPT-4o-mini	39.9	33.4	17.4	22.9	5.0	32.4	11.9	5.2	24.9	17.4	5.1	52.3
Claude-3.5-sonnet	37.6	26.5	16.6	22.4	4.8	19.3	13.9	2.9	17.9	18.1	3.9	60.5
Open-source Models												
Qwen2.5VL-32B	**30.7**	11.7	26.3	17.7	5.4	10.3	20.4	1.9	18.3	19.0	3.6	**53.5**
Qwen2.5VL-7B	17.5	13.6	**27.6**	**17.7**	**5.0**	11.2	**18.9**	**1.9**	**19.4**	**18.3**	**3.4**	43.2
LLaVA-Next-Video-34B	12.4	12.3	0.8	17.4	0.0	11.8	0.9	0.0	6.3	9.2	0.0	48.6
LLaVA-OneVision-7B	12.3	**18.0**	4.6	14.7	0.0	17.0	5.6	0.1	10.8	10.2	0.1	28.3
InternVL2-8B	11.6	7.4	14.5	8.3	1.2	9.5	9.1	0.5	12.0	8.7	0.9	40.2
LLaVA-Next-Video-7B	11.7	17.2	0.0	4.4	0.0	15.9	0.0	0.0	8.0	2.2	0.0	27.3
LongVA-7B	9.2	14.7	2.8	5.0	0.0	10.3	0.7	0.0	6.6	2.8	0.0	49.7
Aria-28B	7.0	12.1	19.0	14.8	0.6	**18.7**	18.1	0.1	18.9	16.5	0.4	43.0
LongVU-7B	3.3	12.2	3.2	1.3	0.3	4.4	2.1	0.5	3.8	1.7	0.4	34.2
Random Chance	9.8	9.8	9.8	9.8	1.0	9.8	9.8	1.0	9.8	9.8	1.0	25.0

2-needle questions. For human evaluation, we adopt majority voting among the five annotators for Yes/No evaluations, and use the average for numerical evaluations.

Tab. 2 shows the evaluation results. The generated questions receive near-perfect scores on all metrics. Reassuringly, all random baselines score near the lowest possible, which is 0% for shared existence of bridge entities and correctness of vague references, and 1 for factual correctness of questions. These results indicate that the automatically generated questions have high quality and the evaluation processes are valid. The prompts for question evaluation are in Appendix Sec. F.

4 Evaluation of VLMs

4.1 Evaluation Setup

We evaluate a total of 13 advanced VLMs, consisting of 4 proprietary models and 9 open-source models. Proprietary models include ChatGPT-4o (gpt-4o-2024-08-06) [OpenAI, 2024a], ChatGPT-4o-mini (gpt-4o-mini-2024-07-18) [OpenAI, 2024b], Gemini-1.5-pro (gemini-1.5-pro-002, Sep 2024) [Team et al., 2024], and Claude-3.5-sonnet (claude-3-5-sonnet-20241022) [Anthropic, 2024]. Open-source models include LLaVA-Next-Video-7B [Zhang et al., 2024b], LLaVA-Next-Video-34B [Zhang et al., 2024b], LLaVA-OneVision-7B [Li et al., 2024a], LongVU-7B [Shen et al., 2024], Aria-28B [Li et al., 2024b], Qwen2.5VL-7B/32B [Bai et al., 2025], LongVA-7B [Zhang et al., 2024a] and InternVL2-8B [Chen et al., 2024]. To establish a human baseline, we employ two annotators to answer 316 randomly selected questions (including causal 1-needle, 2-needle visual grounding and image description). Details about the input content, the chat template of VLMs can be found in Appendix Sec. C.

Most models we tested do not natively support inputs containing multiple video clips. As a workaround, we uniformly sample five frames from each clip and stack them vertically as one input image. This allows us to stay within the image input limit, which is 32 images for most models. Compared to other possible techniques we experimented with, this trick yields the best performance (see Appendix Sec. D).

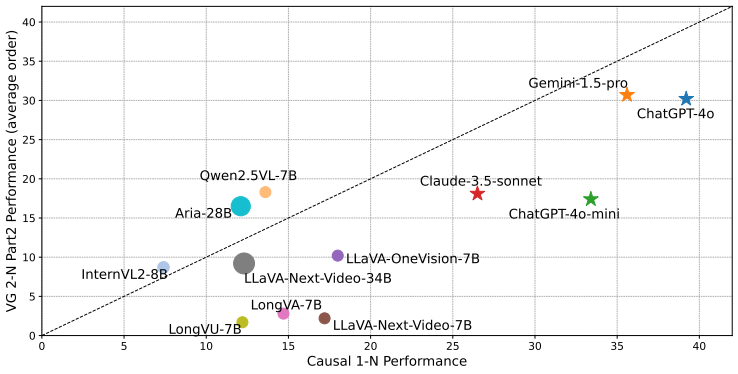

Figure 3: Comparison of performance of models on Part 2 of VG 2-needle questions (average order) and causal 1-needle questions. This is a fair comparison since both ask to retrieve the cause event. The dashed diagonal line represents equal performance across the two question types. Most points fall below the line, indicating that models are poorer at VG 2-needle questions than causal 1-needle questions. The size of the dots indicates the model size. The stars indicate proprietary models.

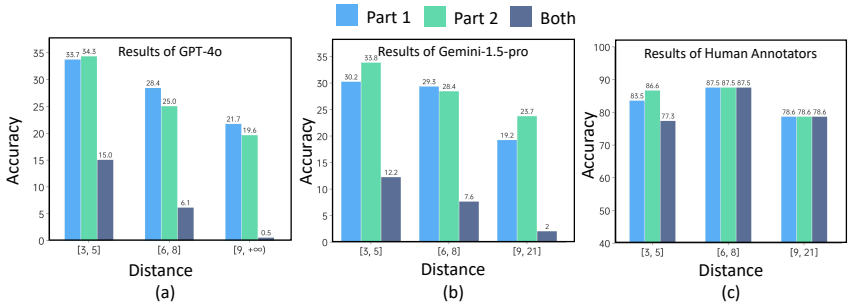

Figure 4: Performance on VG 2-needle questions as the distance between the two needle grows. We report the average-order performance for models and the forward-order performance for human annotators. The model performance declines whereas human performance stays mostly unchanged.

To prevent the models from using sentence indices as shortcuts, we input the full narration along with only a subset of the corresponding video clips. The narration and video clips are provided separately. We avoid aligning video clips with sentences in the input to prevent the models from relying on text to locate video clips. We input all video clips from the cause clip to the effect clip, as well as a total of five clip before and after this span; the numbers of clip before and after are random. This setup prevents the cause clip from always being the first clip and the effect clip always the last, eliminating shortcut learning based on location.

4.2 Main Results

We present the performance of various VLMs on the noncausal 1-needle questions, causal 1-needle questions, Visual Grounding (VG) 2-needle questions, and Image Description (ID) 2-needle questions in Tab. 3. We report the accuracy on each type of questions. For VG 2-needle questions, we separately compute the accuracy for Part 1, Part 2, and both parts answered correctly. In addition, we evaluate each model using both the original (forward) and reversed video clip order, and report the average of the two orders as the final result on VG 2-needle questions. This design is motivated by positional bias of the models, discussed in Sec. 4.3.

Causal Questions Are More Challenging Than Noncausal Questions. From Tab. 3, we observe that most models perform substantially better on noncausal 1-needle questions than on causal ones. For instance, ChatGPT-4o and Qwen2.5-VL-32B achieve results that are 17.6% and 19.0% higher, respectively. This gap suggests that causal reasoning still poses a significant hurdle for current state-of-the-art models on long-video understanding.

Two-needle Questions Are More Challenging Than One-needle. We compare the performance on Part 2 of VG 2-needle questions (average order) and causal 1-needle questions, since both ask for the retrieval of the cause event and constitute a fair comparison. Most models perform worse on the 2-needle questions than the 1-needle questions. Fig. 3 visualizes this pattern. This highlights the deficiency of using only 1-needle questions to evaluate long-video understanding.

Open-Source Models Exhibit Weaker World Modeling Ability. According to Tab. 3, open-source models generally perform worse than proprietary models. Specifically, proprietary models surpass all open-source models on both 1-needle and ID 2-needle questions. For VG 2-needle questions (Avg), all open-source models, except for Qwen2.5VL-7B and Aria-28B, perform at or below random levels, falling behind proprietary models. This gap may be due to insufficient human-behavior training data during the pretraining of open-source VLMs.

Performance Decreases with Increasing Needle Distance. In Fig. 4 (a) and (b), we present the VG 2-needle question performance (average order) for ChatGPT-4o and Gemini-1.5-pro. Performance is measured on questions with varying distances between the cause and effect events. The results show a clear decline as distance increases, especially when the evaluation aggregates answers from both parts. This phenomenon indicates that the distance between multiple needles has a significant impact when joint understanding of these needles is required. More details are provided in Appendix Sec.D.

A possible confounder is the strength of causal relations[2]. That is, long distances between needles may be correlated with weak causal relationships. It may be the weak causality, rather than long distance, that causes performance decrease. To verify this conjecture, we test if humans can identify the cause clip over long distances. Two annotators answered 113 visual grounding 2-needle questions and the results with different needle distances are detailed in Fig. 4 (c). We can observe that for 24 samples with a needle distance between 6 and 8 clips, the human performance is 87.5% (Part 1), 87.5% (Part 2) and 87.5% (Both). This is even better than the performance on 97 samples with a distance between 3 and 5 clips: 83.5% (Part 1), 86.6% (Part 2) and 77.3% (Both). These results demonstrate that humans can easily undestand the causal relations. Thus, we argue that the model performance degradation mainly stems from the long distances between needles.

4.3 Pathological Behaviors of VLMs

Positional Bias. As shown in Tab. 3, the accuracy of models responses varies depending on the positions of cause and effect clips. In VG 2-needle questions, Part 1 is grounded in the effect video clip, while Part 2 is grounded in the cause clip. In the forward-ordered clip sequence, the cause clip appears earlier, and all proprietary models show higher accuracy in Part 2 compared to Part 1; for instance, the accuracy of ChatGPT-4o exhibits an accuracy increase of 22.5%. This result appears paradoxical, since from an information-processing perspective answering the Part 2 question requires an understanding of the video clip required for the Part 1 question. Why are models more accurate when locating the cause clip required by Part 2 than Part 1?

To answer this question, we reverse the order of the input video clips (the Reverse column of Tab. 3). In this setting, Part 1 achieves much higher accuracy than Part 2. This suggests that models pay more attention to clips appearing earlier; therefore, visual grounding for the cause clips is easier under normal playback, while reverse playback facilitates the visual grounding of the effect clips. To mitigate the influence of position bias, we evaluate model accuracy of VG 2-needle questions using the average of forward and reverse clip ordering.

Static Output Bias. In Tab. 3, we observe that some open-source VLMs achieve extremely low scores, sometimes zero, on VG 2-needle questions. When examined closely, we observe that some models tend to output the same prediction regardless of the question content. We call this phenomenon the static output bias.

In Fig. 5, we compare the output distributions of well-performing and poorly-performing models on VG 2-needle questions. For example, the ground truth of Part 1 is mainly in clip 5 and beyond, as the effect clip appears after the cause. ChatGPT-4o shows a balanced distribution, aligning with its strong performance in Tab. 3. In contrast, the three open-source models predominantly select clip 1 as the answer, demonstrating severe output bias. We experiment with different prompting strategies

[2]We thank an anonymous reviewer for suggesting this.

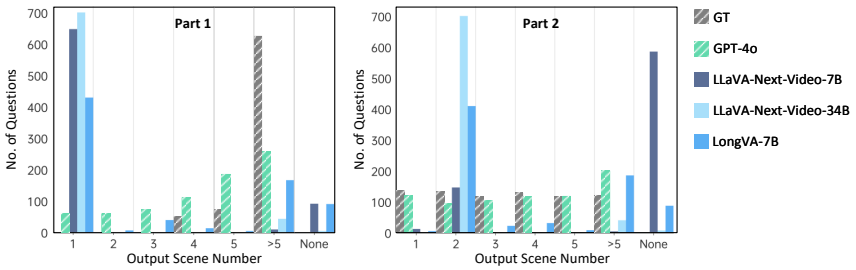

Figure 5: The answer distribution of various models in the forward evaluation of visual grounding 2-needle questions. GT denotes ground truth. None means no clip number is output. Predictions of opensource models are heavily concentrated in a few numbers, exhibiting significant bias.

Table 4: Results of models on different input settings. Numbers in parentheses indicate the absolute increase over random chance. 1-N: 1-needle, 2-N: 2-needle, V: Video, N: Narration, Q: Question. When V is removed in causal 1-N, performance decreases to near chance level, demonstrating minimal textual bias. When V+N are removed in ID 2-N, performances stay above chance level but are still far from saturation.

Models	**Causal 1-N Questions**		**ID 2-N Questions**	
	V+N+Q	N+Q	V+N+Q	Q only
Proprietary Models				
Gemini-1.5-pro	35.6 (+25.7)	12.3 (+2.4)	60.9 (+35.9)	39.0 (+14.0)
Claude-3-5-sonnet	26.5 (+16.7)	10.7 (+0.9)	60.5 (+35.5)	29.2 (+4.2)
ChatGPT-4o	39.2 (+29.4)	3.8 (-6.0)	59.2 (+34.2)	47.5 (+22.5)
Open-source Models				
LongVA-7B	14.7 (+4.9)	11.2 (+1.4)	49.7 (+24.7)	38.4 (+13.4)
InternVL2-8B	7.4 (-2.4)	11.7 (+1.9)	40.2 (+15.2)	25.9 (+0.9)
LLaVA-Next-Video-7B	17.2 (+7.4)	0.3 (-9.5)	27.3 (+2.3)	26.9 (+1.9)

as attempts to fix this bias, but the models consistently produce fixed responses, indicating an inherent flaw in these models. The test prompts and output patterns are in Appendix Sec. F.

4.4 Analysis of Dataset Bias

Textual Bias. Textual bias refers to the extent that the dataset allows models to derive the correct answer from the accompanying text alone. To verify if our visual grounding test format effectively avoids textual bias, we test various models on causal 1-needle questions without providing the video clips. The results are presented in the first two columns of Tab. 4. After removing video input, model performance drops significantly to around random chance (9.8%) or lower. For instance, the performance of ChatGPT-4o drops from 39.2% to 3.8%. These results indicate that our setting effectively mitigates textual bias, ensuring a reliable evaluation of multimodal understanding abilities.

Knowledge Leakage from Pretraining. To prevent performance underestimation caused by the out-of-distribution VG question format, we introduce an image description (ID) format. However, an VLM may have memorized details of movies from pretraining, which can be used to answer these questions. To assess the impact of prior knowledge, we test different models using question-only inputs. The results are shown in the last two columns of Tab. 4. We find that ChatGPT-4o, Gemini-1.5-pro, and LongVA achieve high performance on ID questions with question-only inputs, with +22.5%, +14.0%, and +13.4% over random chance, respectively. This indicates that they can call upon prior knowledge to answer the questions. In contrast, Claude-3.5-sonnet and InternVL2-8B are less affected by prior knowledge. For almost all models tested, the performance with video and narration input is much higher than with question only, suggesting knowledge leakage alone is not sufficient for good performance. There remains substantial room for improvement (around 60–70%) in CAUSAL2NEEDLES, which could benefit from enhanced causal reasoning capability.

Acknowledgments

This research is supported by the RIE2025 Industry Alignment Fund – Industry Collaboration Projects (IAF-ICP) (Award I2301E0026), administered by A*STAR, by Alibaba Group and NTU Singapore through Alibaba-NTU Global e-Sustainability CorpLab (ANGEL), by the Nanyang Associate Professorship, and by the National Research Foundation Fellowship (NRFF13-2021-0006), Singapore.

References

Maryam Amirizaniani, Elias Martin, Maryna Sivachenko, Afra Mashhadi, and Chirag Shah. Do llms exhibit human-like reasoning? evaluating theory of mind in llms for open-ended responses. *arXiv preprint arXiv:2406.05659*, 2024.

Anthropic. Claude 3 model card, 2024. URL `https://www-cdn.anthropic.com/de8ba9b01c9ab7cbabf5c33b80b7bbc618857627/Model_Card_Claude_3.pdf`. Preprint, accessed on March 8, 2025.

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.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde De Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024.

Aditya Chinchure, Sahithya Ravi, Raymond Ng, Vered Shwartz, Boyang Li, and Leonid Sigal. Black swan: Abductive and defeasible video reasoning in unpredictable events. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 24201–24210, 2025.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

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

Pelin Dogan, Boyang Li, Leonid Sigal, and Markus Gross. A neural multi-sequence alignment technique (neumatch). In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 8749–8758, 2018.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36:70293–70332, 2023.

Jay Wright Forrester. Counterintuitive behavior of social systems, 1971. URL `https://en.wikipedia.org/wiki/Mental_model`.

Simon Frieder, Luca Pinchetti, Ryan-Rhys Griffiths, Tommaso Salvatori, Thomas Lukasiewicz, Philipp Petersen, and Julius Berner. Mathematical capabilities of chatgpt. *Advances in neural information processing systems*, 36:27699–27744, 2023.

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.

Yanchen Guan, Haicheng Liao, Zhenning Li, Jia Hu, Runze Yuan, Yunjian Li, Guohui Zhang, and Chengzhong Xu. World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles*, 2024a.

Yanchen Guan, Haicheng Liao, Zhenning Li, Jia Hu, Runze Yuan, Yunjian Li, Guohui Zhang, and Chengzhong Xu. World models for autonomous driving: An initial survey. *IEEE Transactions on Intelligent Vehicles*, 2024b.

Wes Gurnee and Max Tegmark. Language models represent space and time. *arXiv preprint arXiv:2310.02207*, 2023.

David Ha and Jürgen Schmidhuber. World models. *arXiv preprint arXiv:1803.10122*, 2018.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Yi Hu, Xiaojuan Tang, Haotong Yang, and Muhan Zhang. Case-based or rule-based: how do transformers do the math? In *Proceedings of the 41st International Conference on Machine Learning*, pages 19438–19474, 2024.

Gregory Kamradt. Needle in a haystack - pressure testing LLMs., 2023. URL `https://github.com/gkamradt/LLMTest_NeedleInAHaystack/tree/main`.

Bingyi Kang, Yang Yue, Rui Lu, Zhijie Lin, Yang Zhao, Kaixin Wang, Gao Huang, and Jiashi Feng. How far is video generation from world model: A physical law perspective. *arXiv preprint arXiv:2411.02385*, 2024.

Dohwan Ko, Ji Soo Lee, Woo-Young Kang, Byungseok Roh, and Hyunwoo J Kim. Large language models are temporal and causal reasoners for video question answering. In *The 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Mosh Levy, Alon Jacoby, and Yoav Goldberg. Same task, more tokens: the impact of input length on the reasoning performance of large language models. *arXiv preprint arXiv:2402.14848*, 2024.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024a.

Dongxu Li, Yudong Liu, Haoning Wu, Yue Wang, Zhiqi Shen, Bowen Qu, Xinyao Niu, Fan Zhou, Chengen Huang, Yanpeng Li, et al. Aria: An open multimodal native mixture-of-experts model. *arXiv preprint arXiv:2410.05993*, 2024b.

Haoxin Li, Yingchen Yu, Qilong Wu, Hanwang Zhang, Song Bai, and Boyang Li. Learning to animate images from a few videos to portray delicate human actions. *arXiv Preprint 2503.00276*, 2025.

Jiaqi Li, Mengmeng Wang, Zilong Zheng, and Muhan Zhang. Loogle: Can long-context language models understand long contexts? *arXiv preprint arXiv:2311.04939*, 2023.

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, 2024c.

Hao Liu, Wilson Yan, Matei Zaharia, and Pieter Abbeel. World model on million-length video and language with blockwise ringattention. *International Conference on Learning Representations*, 2025.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173, 2024.

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.

Iman Mirzadeh, Keivan Alizadeh, Hooman Shahrokhi, Oncel Tuzel, Samy Bengio, and Mehrdad Farajtabar. Gsm-symbolic: Understanding the limitations of mathematical reasoning in large language models. *arXiv preprint arXiv:2410.05229*, 2024.

Saman Motamed, Laura Culp, Kevin Swersky, Priyank Jaini, and Robert Geirhos. Do generative video models learn physical principles from watching videos? *arXiv preprint arXiv:2501.09038*, 2025.

OpenAI. Chatgpt-4o (aug 6 version), 2024a. URL `https://openai.com/index/hello-gpt-4o/`. Large language model.

OpenAI. Chatgpt-4o-mini (july 18 version), 2024b. URL `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`. Large language model.

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.

Aaditya K Singh, Muhammed Yusuf Kocyigit, Andrew Poulton, David Esiobu, Maria Lomeli, Gergely Szilvasy, and Dieuwke Hupkes. Evaluation data contamination in llms: how do we measure it and (when) does it matter? *arXiv preprint arXiv:2411.03923*, 2024.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

Yidan Sun, Qin Chao, Yangfeng Ji, and Boyang Li. Synopses of movie narratives: a video-language dataset for story understanding. *arXiv preprint arXiv:2203.05711*, 2022.

Yidan Sun, Qin Chao, and Boyang Li. Event causality is key to computational story understanding. In *The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2024. URL `https://arxiv.org/abs/2311.09648`.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Kiran Vodrahalli, Santiago Ontanon, Nilesh Tripuraneni, Kelvin Xu, Sanil Jain, Rakesh Shivanna, Jeffrey Hui, Nishanth Dikkala, Mehran Kazemi, Bahare Fatemi, et al. Michelangelo: Long context evaluations beyond haystacks via latent structure queries. *arXiv preprint arXiv:2409.12640*, 2024.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Xiaotao Gu, Shiyu Huang, Bin Xu, Yuxiao Dong, et al. Lvbench: An extreme long video understanding benchmark. *arXiv preprint arXiv:2406.08035*, 2024a.

Weiyun Wang, Shuibo Zhang, Yiming Ren, Yuchen Duan, Tiantong Li, Shuo Liu, Mengkang Hu, Zhe Chen, Kaipeng Zhang, Lewei Lu, et al. Needle in a multimodal haystack. *arXiv preprint arXiv:2406.07230*, 2024b.

Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Linjie Li, Jena D Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Chandu, et al. The generative ai paradox:" what it can create, it may not understand". *arXiv preprint arXiv:2311.00059*, 2023.

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.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. Reasoning or reciting? exploring the capabilities and limitations of language models through counterfactual tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 1819–1862, 2024.

Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024.

Zi Yang. Retrieval or holistic understanding? dolce: Differentiate our long context evaluation tasks. *arXiv preprint arXiv:2409.06338*, 2024.

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*, 2024a.

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*, 2024b.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36:46595–46623, 2023.

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.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: In the abstract, we clearly state our primary contribution, which is the design of a novel set of benchmark questions specifically created to evaluate models' abilities in long-context video understanding. In the introduction, we highlight two key limitations of existing video understanding datasets to underscore the necessity and effectiveness of our proposed benchmark. We further summarize the performance of contemporary video large language models and identify their strengths and weaknesses. Thus, these claims represent the primary contributions of our paper, and these contributions are also summarized explicitly in the introduction.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: In Appendix Section E.

   Guidelines:

   - The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper does not include theoretical results.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: For question generation, the procedure are introduced in Section 3. For model evaluation, We comprehensively introduced the models used in our experiments as well as the experimental procedures in Section 4. Additionally, we have publicly released both the code and benchmark dataset to facilitate reproduction and verification of our main experimental results.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: Following the guidelines, we have fully disclosed all information necessary to reproduce our main experimental results. We have publicly released both our benchmark dataset (available at: `https://huggingface.co/datasets/causal2needles/Causal2Needles`) and the code used for evaluating model performance (available at: `https://github.com/jdsannchao/Causal2Needles`). For proprietary models, we set the `temperature` parameter to 0 to minimize randomness in generation. For open-source models, we set `do_sample` to `False` to ensure deterministic outputs. These settings allow our reported results to be reliably reproduced.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

   Answer: [Yes]

   Justification: In Section 4.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: To minimize randomness and ensure reproducibility, all models were evaluated using fixed decoding settings. Specifically, we set `temperature` to 0 for proprietary models and `do_sample` to `False` for open-source models. As a result, model outputs are consistent across runs, and standard statistical significance tests or error bars are not applicable in this setting.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: In Appendix Section C.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Yes, the research conducted in this paper fully conforms to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.

- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [NA]

    Justification: Our method primarily influences the evaluation of large video language models and may encourage the development of stronger video understanding capabilities. It does not have any direct societal impact, nor does it pose any foreseeable negative consequences.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: No additional safeguards were required, as the dataset is used strictly for its intended academic purpose of evaluating model performance. The datasets used in our study are all publicly available, originally sourced from YouTube videos. We do not introduce any new high-risk data or models. Our benchmark is constructed purely for academic evaluation purposes and does not contain sensitive, private, or potentially misusable content.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

    Justification: All referenced datasets have been properly cited.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: All new assets introduced in the paper, including our benchmark dataset and evaluation scripts, are well documented.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: There is no crowd sourcing experiments.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [Yes]

    Justification: This paper does not involve any potential risks. For human evaluation of the dataset, we obtained IRB approvals.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [Yes]

    Justification: The use of LLMs is a central component of our methodology. LLMs are involved in every step of our question generation pipeline. We explicitly describe the role of LLMs in Section 3.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.

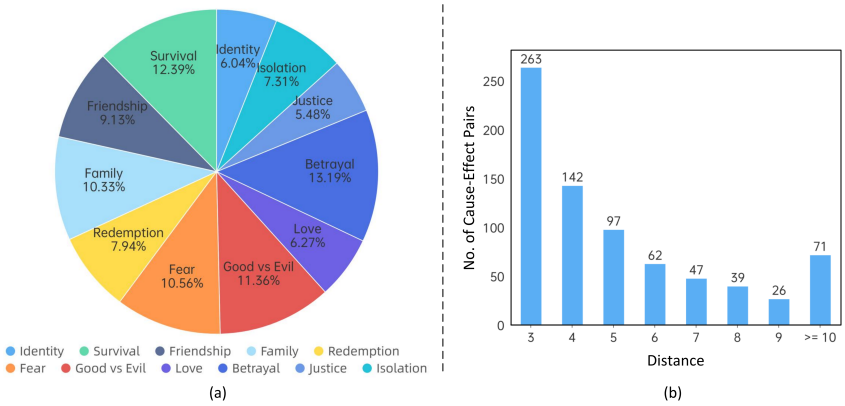

Figure 6: (a) Distribution of movie themes in CAUSAL2NEEDLES. (b) Distribution of distance between cause and effect in CAUSAL2NEEDLES

A Dataset Statistics

The videos in CAUSAL2NEEDLES are collected from the SyMoN [Sun et al., 2022] and YMS [Dogan et al., 2018] datasets, comprising movie recap videos and their corresponding narrations (98 movies from SyMoN and 94 from YMS). After integration, the videos in CAUSAL2NEEDLES have an average duration of 438 seconds, with 58.6 clips and 1,267.2 words per movie. The movie themes in our dataset are highly diverse. To quantify this diversity, we first use ChatGPT-4o to assign multi-labels to each narrative based on 11 predefined thematic categories, and then compute the frequency of each predicted label. The distribution is shown in Fig. 6 (a). As detailed in Sec. 3, we extract 747 cause-effect pairs from the narrations and construct 2-needle questions accordingly. We further analyze the distribution of their distances, defined as the number of clips between the cause and effect events, shown in Fig. 6 (b).

B Event Graph Merging

To obtain comprehensive and long-range causal relationships, we merge the extracted local and global event graphs. The merging process simply takes the union of the causal relations from different event graphs. For example, suppose the output of Graph A is [(1, 5), (3, 10), (21, 27)]. Here, (1, 5) denotes a causal relation from event 1 to event 5. Further suppose the output of Graph B is [(1, 5), (3, 10), (6, 11)]. The merged result is the union [(1, 5), (3, 10), (6, 11), (21, 27)].

C Evaluation Details

Computing Resources. We conduct experiments using four NVIDIA GeForce RTX A6000 GPUs for all open-source models. For proprietary models, we obtain results through their publicly available APIs. The inference batch size is set to 1.

Details of VLMs. Current video understanding VLMs primarily rely on converting videos into sequences of images for processing. For example, LLaVA-Next-Video [Zhang et al., 2024b] samples 32 frames from a video, encodes them into visual tokens using a Vision Transformer (ViT) [Dosovitskiy et al., 2020], and feeds these tokens, along with textual instructions, into Vicuna-1.5 [Zheng et al., 2023] to generate responses. Interleaved VLMs [Chen et al., 2024, Li et al., 2024b] extend input formats to support text-image sequences.

Proprietary models such as Gemini-Pro [Team et al., 2024], GPT-4o [OpenAI, 2024a], and Claude-3.5 [Anthropic, 2024] benefit from substantially greater computational resources, enabling them to process millions of image tokens—achieving dense visual understanding at up to 2 frames per second. These models also support interleaved text-image inputs, offering stronger temporal and contextual reasoning capabilities over video content.

Next, we show the input content and format we used when evaluating different models.

Proprietary LLMs. For GPT-4o, GPT-4o-mini, Gemini-1.5-Pro-002, and Claude-Sonnet, the input consists of: (1) A series of movie scene images, including the scene containing the correct answer. Each scene is represented by five video frames stitched together vertically. (2) The task instruction, which specifies the task type—either visual grounding or image description. In the visual grounding task, the model is instructed to identify the correct scene number, whereas in the image description task, the model is asked to select the correct option. (3) The movie's narration background. (4) The question. If the task is image description, the input will also include textual answer options. Thus, each input is structured as shown in the list below:

```
<image> <image> <image> ...
<text: Test Instruction>
<text: Movie Narration>
<text: Questions>
```

Open-source VLMs. For LLaVA-Next-Video-7B, LLaVA-Next-Video-34B, LLaVA-OneVision-7B, LongVU-7B, and LongVA-7B, the input formats are the same as those used for the proprietary models. However, these models utilize different chat templates during generation. We follow the examples presented on each model's GitHub webpage or Hugging Face repository.
LLaVA-Next-Video-7B: `https://huggingface.co/llava-hf/LLaVA-NeXT-Video-7B-hf`
LLaVA-Next-Video-34B: `https://huggingface.co/llava-hf/LLaVA-NeXT-Video-34B-hf`
LLaVA-OV-7B: `https://huggingface.co/llava-hf/llava-onevision-qwen2-7b-ov-hf`
LongVU-7B: `https://huggingface.co/Vision-CAIR/LongVU_Qwen2_7B`
LongVA-7B: `https://github.com/EvolvingLMMs-Lab/LongVA`

The remaining three are VLMs pretrained on interleaved image-text data. To match their input format, we insert scene numbers between video segments. However, we do not insert any text between scenes, as this could provide shortcuts by overly aligning with the narration, rather than requiring true visual understanding.

```
scene1<image>scene2 <image>scene3<image>
<text: Test Prompt>
<text: Movie Narration>
<text: Questions>
```

The chat template is collected from:
InternVL2: `https://huggingface.co/OpenGVLab/InternVL2-8B`
Aria-28B: `https://huggingface.co/rhymes-ai/Aria`
Qwen2.5VL-7B: `https://huggingface.co/Qwen/Qwen2.5-VL-7B-Instruct`

Random Chance Calculation. For a 1-needle question, the model is required to select the correct scene from a sequence of given scenes (including extra padding scenes). Assuming the total number of input scenes is N, the random chance probability is $1/N$ (reported as a percentage). For a visual grounding 2-needle question, where the model must make two separate selections, the random chance of answering both questions correctly is $1/N^2$.

Since the distance between the cause and effect scenes varies, the total number of input scenes N differs for each question. Therefore, we compute the average input length $\bar{N}$ over each type of questions and determine the random chance probability. For 1-needle questions, the random chance probability is 9.8%. For visual grounding 2-needle questions, the probability of correctly identifying a single part is 9.8%, while the probability of getting both correct is only 1.0%. For image description 2-needle questions with four options, the random chance is 25.0%.

D Additional Experiments

In this section, we provide a deeper analysis of several aspects: (1) the influence of the distance between cause and effect events; (2) textual grounding versus visual grounding; (3) the uniqueness of answers to Part 2 questions; (4) existence of counterfactual answers in ID 2-needle questions; and (5) non-stacked versus stacked image input settings.

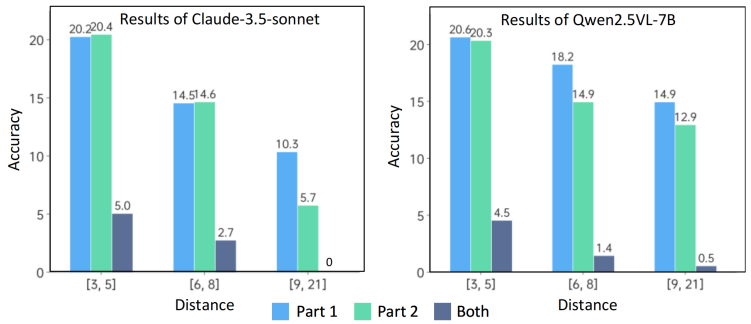

Figure 7: The 2-needle visual grounding performance of various models on questions with different needle distances.

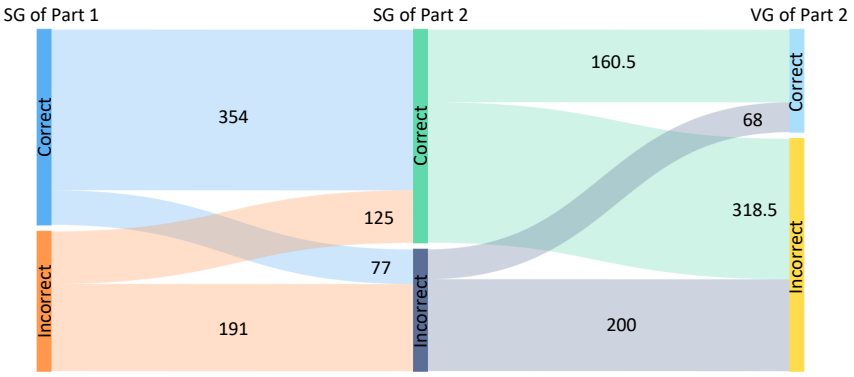

Figure 8: Performance of ChatGPT-4o on the sentence grounding (SG) task and the visual grounding (VG) task on the 2-needle questions. In the SG tasks, the sentences are always arranged in their natural order. In the VG task, we take the average of the forward and reverse orderings.

The Influence of Needle Distance. Fig. 4 (b) in Sec. 4.2 shows that the 2-needle visual grounding performance decreases as the distance between the cause and effect events increases, based on evaluations with two proprietary models. Here, we provide additional results in Fig. 7. These results indicate that needle distance significantly impacts model performance in multi-needle questions.

Textual vs. Visual Grounding. To better understand the mechanisms underlying the models' performance on VG 2-needle questions, we conduct additional experiments to analyze visual and textual grounding. We introduce a new task called Sentence Grounding (SG), which requires the model to select the sentence that contains the answer from the story narration text, which are part of the input of CAUSAL2NEEDLES. We compare the SG performance of GPT-4o on Part 1 and Part 2 questions, and summarize the results in Fig. 8. We make the following observations.

First, the results show that the Part 2 question relies on joint understanding of both the cause and the effect sentences. We compute the conditional probability that GPT-4o answers SG Part 2 correctly given that it answers SG Part 1 correctly, $\mathrm{P}(\text{SG-Part2} = \text{correct} \mid \text{SG-Part1} = \text{correct}) = 82.1\%$. In contrast, the conditional probability that GPT-4o answers SG Part 2 correctly given that it answers SG Part 1 incorrectly, $\mathrm{P}(\text{SG-Part2} = \text{correct} \mid \text{SG-Part1} = \text{incorrect}) = 39.6\%$. The large gap between the two conditional probabilities demonstrates that performance on SG Part 2 significantly depends on SG Part 2. This suggests the existence of a sequential dependency between the two parts. The finding demonstrates that, at least in the textual modality, the two-needle questions benefit from the joint understanding of both needles.

Second, the dependence of VG performance on SG performance is limited. We compute the conditional probability that GPT-4o answers VG Part 2 correctly given that it answers SG Part 2 correctly, $\mathrm{P}(\text{VG-Part2} = \text{correct} \mid \text{SG-Part2} = \text{correct}) = 33.5\%$. In contrast, the conditional probability that the model answers VG Part 2 correctly given that it answers SG Part 2 incorrectly is $\mathrm{P}(\text{VG-Part2} = \text{correct} \mid \text{SG-Part2} = \text{incorrect}) = 25.4\%$. The difference between the two

Table 5: Comparison of results between stacked image inputs and non-stacked image inputs.

Models	Input Setting	VG 2-Needle Questions								
		Forward			Reverse			Avg		
		Part 1	Part 2	Both	Part 1	Part 2	Both	Part 1	Part 2	Both
ChatGPT-4o	non-stacked	5.0	29.4	1.4	29.7	4.1	0.9	17.4	16.8	1.2
	stacked	**16.7**	**39.2**	**9.4**	**45.4**	**21.2**	**13.4**	**31.1**	**30.2**	**11.4**
Gemini-1.5-pro	non-stacked	5.4	37.3	2.9	26.3	17.5	6.1	15.8	27.4	4.5
	stacked	**21.0**	**40.0**	**10.2**	**35.7**	**21.4**	**8.4**	**28.4**	**30.7**	**9.3**
Qwen2.5VL-7B	non-stacked	2.1	11.0	0.0	**13.8**	4.4	0.4	8.0	7.7	0.2
	stacked	**27.6**	**17.7**	**5.0**	11.2	**18.9**	**1.9**	**19.4**	**18.3**	**3.4**

probabilities is small. This shows the correctness of SG Part 2 only makes minor contribution to VG Part 2. That is, even if the model is wrong on SG Part 2, it still has a decent chance of being correct on VG Part 2, relative to the alternative situation that the model is correct on SG Part 2. Hypothetically, a main reason for this phenomenon is that the model utilizes a separate mechanism for visual reasoning, which operates more or less independently of reasoning over the textual modality.

Lastly, on SG Part 2, GPT-4o achieves 63.7% accuracy, much higher than the 30.3% on VG Part 2, suggesting the model is better at textual than visual reasoning. This large gap indicates substantial room for improvement in visual comprehension of VLMs.

Answer Uniqueness of Part 2 Questions. We conduct an experiment to verify if Part 2 questions have unique answers. In causal relationship extraction, an event graph may contain multiple nodes pointing to the same node, indicating that a single event can have multiple causes. This may introduce multiple ground truth answers to Part 2 questions, as they require retrieving the cause of an event.

Our experiment is designed as follows: among the 747 causal relationships of 2-needle questions, we identify 99 instances where the effect sentence is not uniquely paired with a single cause sentence. For each effect sentence in them, we feed the corresponding Part 2 question and cause sentences (on average, 2.02 candidate causes per effect) to ChatGPT-4.1. Then, it is asked to select the cause sentence that contains the answer to the Part 2 question. The results show that out of the 99 such cases, ChatGPT-4.1 selects the wrong cause sentence in only one instance. This indicates that the answer to Part 2 question is mostly unique.

Existence of Counterfactual Answers in ID 2-Needle Questions. A counterfactual answer refers to an answer that is supported apparently by the input video and can be ruled out only by causal reasoning. If the video contains only the correct answer and no counterfactual answers, a neural network can easily rule out wrong answers without locating the required video clip.

For example, in Fig. 2, the image description question is "What tragedy was commemorated by the memorial service in Metropolis? When the tragedy that causes the memorial service occurs, what color were the energy blasts emanating from the creature? A. Yellow B. Blue C. Red D. White." The correct answer is "Yellow", which appears in the cause clip. "Blue" is a competing answer as a blue light is shown in another battle clip at 2'37 of the video. Due to the presence of both answers in the input video, to correctly answer the question, the model must identify the cause clip through causal reasoning.

We adopt two question design methods to ensure the existence of counterfactual answers in image description questions. First, we are deliberately ambiguous about the reference in the question wording, as discussed in Sec. 3.3. Second, we encourage the LLM to search for visually similar but potentially confusing content from adjacent clips, using the instructions in Appendix Fig. 14.

To demonstrate the existence of counterfactual answers, we conduct an experiment with the following setting: (1) We remove the cause clip from the input and any information that helps to locate the cause clip from the question, allowing the model to answer the question using other clips. For example, the aforementioned question becomes " What color were the energy blasts emanating from the creature?"; (2) We replace the ground truth answer with "None is correct," so if there are no counterfactual answers, the model should choose this option. In experiments with ChatGPT-4o, the model selects

"None is correct" for only 17.8% of questions. This suggests that most image description questions in CAUSAL2NEEDLES have counterfactual answers.

Non-stacked vs. Stacked Image Input. We explain the rationale for the design choice of inputting each video clip by stacking the video frames into a single image. We represent a video clip using five uniformly sampled frames. After that, we stack these frames from top to bottom as a single image as input to the VLMs. There are two reasons for the stacked approach. First, most open-source models have a limited window size and can only process 32 frames at a time. Therefore, without stacking, processing multiple video clips would quickly exceed the window size. Second, for proprietary models that can accept more frames as input, stacking the frames can help the model better capture and comprehend the story within a video clip. As verification, we conduct additional experiments on VG 2-needle questions using ChatGPT-4o, Gemini-1.5-pro and Qwen2.5VL-7B with non-stacked frames as input. For a fair comparison, we instruct the model to output the frame number containing the answer and calculate the clip number by dividing the frame number by five, as every five frames belong to a single clip. The results are detailed in Tab. 5. We can observe that stacked frames improve performance significantly, indicating that the stacked method enhances visual understanding.

E Limitations

While our benchmark addresses several limitations of prior work, such as the lack of multi-needle evaluation on long-context video understanding and the understanding of human behaviors, there remain areas for future improvement.

First, our videos are primarily sourced from movie clips, which provide a rich set of event causal relations. Compared to real life, movies tend to over-emphasize uncommon events, such as serendipity, intricate conspiracy, or dramatic conflicts. These events could serve as effective out-of-distribution tests for models primarily trained on common events such as ego-centric videos. However, they should not be taken as necessarily representative of real-life events. Second, our current evaluation focuses on VLMs. Many of them, particularly open-source ones, do not yet support audio inputs. Accordingly, we exclude audio from our inputs, but future extensions of the benchmark will incorporate audio to enable evaluation of more comprehensive multimodal understanding.

F Generation and Evaluation Prompts

We present prompting templates for causal-relationship extraction (Fig. 9), 1-needle question generation (Fig. 10), visual-grounding 2-needle question generation (Fig. 12), and image-description 2-needle question generation (Fig. 14). Visual-grounding 2-needle questions and 1-needle questions are generated with GPT-o1 (openai-o1-preview-2024-09-12), with sentence rephrasing performed by Gemini-1.5-Pro-002 as part of the visual-grounding 2-needle pipeline; image-description 2-needle questions are generated with Gemini-1.5-Pro-002.

In Fig. 15, we continue using the same model to generate answer choices. At this stage, to increase the difficulty of the options, we input the cause clip as well as the two adjacent frames and prompt the model to consider generating more challenging options based on objects appearing in the remaining scenes.

Causal Relationships Extraction (GPT-o1-preview-2024-09-12)

Here is a list of nodes (events) from a story event graph. We want you to fill in the edges of the event graph with causal connections between nodes. An event graph contains nodes and edges. Each node represents an event, and each edge represents the causal connection between two events.

Example Input:
Node 0: When Dan goes to school in the morning, he has to take the bus.
Node 1: One day Dan was running late, and missed the bus to school.
Node 2: Dan called his friend Pete, and asked for a ride to school.
Node 3: Pete gave Dan a ride to school, but Dan was late for his first class.
Node 4: Luckily Dan wasn't late for any of his other classes that day.

Example Output:
Edge 0: (Node 0 -> Node 1)
Edge 1: (Node 1 -> Node 2)
Edge 2: (Node 2 -> Node 3)
Edge 3: (Node 1 -> Node 3)
Edge 4: (Node 3 -> Node 4)
(continue with another five demonstrations)

Now, it is your turn to construct the event graph for the following event list.
Event List:
Node 0: <S1>
Node 1: <S2>
Node 2: <S3>
Node 3: <S4>
Node 4: <S5>
Output:

Figure 9: The prompt for causal relationships extraction.

1-Needle Question Generation (GPT-o1-preview-2024-09-12)

Caulsal 1-Needle Question

Given a pair of sentences with a causal relationship, create a question about the “Effect" sentence such that the answer to this question is found exclusively in the “Cause" sentence. Ensure that the answer does not appear within the “Effect” sentence itself.

Example Input: Cause: Blaming himself, Harry Hart, codenamed Galahad, delivers a medal for valor to Lee’s widow, Michelle and his young son, Eggsy, saying that if they ever need help, they should call the phone number on the back of the medal. Effect: Arrested for stealing a car, Eggsy calls the number on the medal.

Example Output: Why did Eggsy have a phone number to call when he was arrested for stealing a car? (continue with another four demonstrations)

Input:
Cause: `<Cause Sentence>`
Effect: `<Effect Sentence>`

Output:
. . . . . .

Noncausal 1-Needle Question

Given a sentence, please use the content of it as context to create a question whose answer appears in the sentence. Please output the question and answer in the following format, keeping the answer as concise as possible:
Question: `<Question>`
Answer: `<Answer>`

Input:
Sentence: `<Cause Sentence/Effect Sentence>`

Output:

Figure 10: The prompt for generating 1-needle questions.

Visual Grounding 2-Needle Question Generation (Step 1)
(Gemini-1.5-pro-002)

Step 1: Rephrase the cause and effect sentences
You have a pair of sentences: one indicating a cause and the other describing its effect. Each sentence has a related image that complements the content expressed in the sentence. Your goal is to rephrase both sentences so that the causal relationship between them is explicitly clear. Please ensure that after rephrasing, there is a piece of shared information present in both sentences that establishes the causal relationship between them. Remember that the rephrased sentences should not include any content that is not present in the original text or images.

Example 1:
Input:
Cause: Gordon, who was actually sent to save Rachel, is unable to make it there in time and Rachel dies.
`<insert cause scene here>`
Effect: The Joker is able to locate Dent in a hospital and manipulates him into seeking revenge for Rachel's death.
`<insert effect scene here>`
Output:

Cause: Gordon, tasked with rescuing Rachel, arrives too late, and she is dead.
Effect: The Joker confronts Dent in a hospital and manipulates him into seeking revenge for Rachel's death against Gordon.

Example 2:
Input:
Cause: The remaining robber reveals himself to be the joker, a crazed supervillain, and escapes the bank in a school bus.
`<insert cause scene here>`
Effect: Several mob leaders hold a conference, which is interrupted by the joker.
`<insert effect scene here>`
Output:

Cause: After the final robber reveals himself to be the joker—an unstable supervillain—and escapes the bank by blending into a line of school buses, the criminal underworld is thrown into disorder.
Effect: To restore order, several prominent mob leaders convene a clandestine meeting, only to have the joker himself crash their conference.

Input:
Cause: `<Cause Sentence>` `<Cause Scene Images>`
Effect: `<Effect Sentence>` `<Effect Scene Images>`

Figure 11: The prompt for generating visual grounding 2-needle questions.

Visual Grounding 2-Needle Question Generation (Step 2)
(GPT-o1-preview-2024-09-12)

Step 2: Generate questions based on bridge entities
You have a pair of sentences: one indicating a cause and the other describing its effect. Your goal is to create a sentence containing two questions. Please complete this goal step by step, as shown in the examples. (Please follow the output format shown in the example.)

Example Input:
Cause: Feeling responsible, Harry Hart, known as Galahad, gives a medal for valor to Lee's widow, Michelle, and his young son, Eggsy, instructing them to call the number on the medal if they ever need help. Effect: When Eggsy is arrested for stealing a car, he calls the number on the medal for help.
Example Output:
Step 1:
Instruction: Identify a piece of shared information that appears in both "Cause" and "Effect" sentences and establishes their causal relationship.
Output: The shared information is 'call the number on the medal'.
Step 2:
Instruction: Create a question that uses the "Effect" sentence as context and shared information as the answer. Rephrase the shared information to be less specific.
Output: First, according to the context, 'call the number on the medal' can be rephrased into 'an approach'. Then, the question is 'What approach did Eggsy take to seek help when he was arrested for stealing a car?'
Step 3:
Instruction: Based on the previous question, create another question using the "Cause" sentence. The answer should be the shared information.
Output: The question is 'Why was he able to adopt such an approach?'
Step 4:
Instruction: Combine the two questions into one sentence.
Output: What approach did Eggsy take to seek help when he was arrested for stealing a car, and why was he able to adopt such an approach?
Input:

Cause: `<Rephrased Cause Sentence>`
Effect: `<Rephrased Effect Sentence>`
Output:

Figure 12: The prompt for generating visual grounding 2-needle questions (continued).

Image Description 2-Needle Question Generation (Step 1-2) (Gemini-1.5-pro-002)

Step 1: You are given a question and its corresponding context. Your task is to concisely identify the following:
The question asks for specific information about an event. Identify this event (which should match the effect event mentioned in the context) and summarize it in a short phrase as **Effect Event**. Then, in the cause sentence, identify the cause event that leads to the effect event found in the first step, and summarize it in a short phrase as **Cause Event**.

Output:
Cause Event: `<cause event>`
Effect Event: `<effect event>`

Input:
Context: `<context>`
Question: `<part1>`

Step 2: You are given a context that can answer a two-part question concerning the same concept (the “bridge entity")—though it may be referred to as a “mission," “event," “approach," or another term in the question. Identify the bridge entity in the two-part question and clarify it based on the context. It often follows “what" or “which".
Example:

Question: What event caused Bryant to order Deckard to retire the replicants, and what is Deckard’s profession that relates to such an event?
Context: Deckard, a Blade Runner tasked with "retiring" replicants, is informed that four—Leon, Roy Batty, Zhora, and Pris—have illegally arrived on Earth. Due to the arrival of the four replicants (Leon, Roy Batty, Zhora, and Pris), Bryant orders Deckard to retire them.

Output:
Bridge Entity: the event
Reference: arrival of the four replicants

Figure 13: The prompt for generating image description 2-needle questions.

Image Description 2-Needle Question Generation (Step 3-4) (Gemini-1.5-pro-002)

Step 3: You are given five images. Your task is to generate a specific visual question with an answer that fits the following format:
"When the event that causes `<effect event>` occurs, `<visual question>`? Answer: `<a short answer to the visual question>`."
Guidelines:
- Ensure the visual question is specific—it should focus on aspects such as environment, object attributes, facial expressions, or clothing.
- Avoid vague questions like "What is the significant object?"
- Avoid well-known facts such as "What is the color of Superman's suit?"
- Avoid overly obvious questions like "What is the doctor wearing?" if the answer is trivial.
- Instead, ask detailed and challenging questions like:
"What color is the jacket he is wearing?"
"How is the character's facial expression?"
"What objects are placed on the table?"
"What is the lighting condition in the background?"

Output: When the event that causes `<EFFECT EVENT>` occurs, ...

Step 4: You are given a question that asks about Event B, which leads to Event A, meaning it contains references to two events.
Your task is to rewrite the question so that it reads naturally while ensuring the following modifications: - Replace any mention of "`<REFERENCE>`" in the question with its corresponding bridge entity: "`<BRIDGEENTITY>`". - Ensure the rewritten question is grammatically correct and sounds natural. - Maintain the focus on Event B (do not rewrite the question to ask only about Event A).

Input: Question: `<question>`

Output: Rewrite question:

Figure 14: The prompt for generating image description 2-needle questions (continued).

Image Description 2-Needle Question Options Generation (Gemini-1.5-pro-002)

You will be given three images along with a question that has a correct answer. Your task is to generate three additional challenging answer choices.

How to Generate Challenging Options:
- If the correct answer involves attributes like color, pattern, shape, or emotion, create alternative choices with different values within the same category.
- Increase difficulty by using existing objects that appear in the images.

Output format: `<Question>`
A. `<Option A>` B. `<Option B>` C. `<Option C>` D. `<Option D>`
Correct Answer: `<Letter>`

Input: Question and its correct answer: `<Question>`

Output:

Figure 15: The prompt for generating options for image description 2-needle questions.

Question Quality Evaluation Prompts We present the prompts used for evaluating the question quality (Bridge Entity, Semantic Coherence, and Readability) in Fig. 16

Question Quality Validation Prompts

Factor 1
You are given two causally related sentences and a bridge entity.
You need to finish the following task: determine whether this bridge entity is a piece of shared information that establishes the causal relationship between the two sentences. You should answer 'yes' or 'no'.

Cause: `<Cause>`
Effect: `<Effect>`
Bridge Entity: `<Bridge Entity>`

.....

Factor 2

You are given a bridge entity and its vague reference.
You need to finish the following task: determine whether this vague reference is more ambiguous than the original bridge entity but still preserves its core meaning. You should answer 'yes' or 'no'.

Bridge Entity: `<Bridge Entity>`
Vague Reference: `<Vague Reference>`

.....

Factor 3

You are given two causally related sentences and a question.
Please evaluate the *semantic coherence* of the question w.r.t the sentences — that is, whether the question matches the facts and events described in both sentences and does not introduce unrelated or contradictory content.

You can only output a score from 1 to 5, where 5 = excellent semantic coherence.

Cause: `<Cause>`
Effect: `<Effect>`
Question: `<Question>`

.....

Factor 4

You will be given a question. Please rate its readability on a scale from 1 to 5, where 5 means very readable and 1 means very not readable.
Consider the following criteria: Naturalness; Grammar; Clarity. You can only output one number.

Question: `<Question>`
Score:

Figure 16: The prompt used for evaluating the quality of generated questions.

Cause-Effect Uniqueness Test Prompts We present the prompts used to test the uniqueness of the generated questions with respect to multiple candidate cause sentences in Fig. 17

Model Performance Evaluation Prompts We present the task instruction prompts used for testing the models in Fig. 18
For visual grounding 2-needle questions, there are two modes: Forward playback and Reverse playback. This distinction is explicitly stated in the instruction prompt.

Various Instruction Templates and Model Output Examples In the Sec. 4.3 of the main text, we observed that many open-source models exhibit severe output bias. Specifically, when testing the visual grounding 2-needle questions, these models tend to produce the same answer ("Scene 1 for part 1 and Scene 2 for part 2.") across different questions. To investigate whether this bias was caused by the prompt, we tested the models with different variations of the prompt. However, we still found that they continued to generate highly consistent responses. Below, in Fig.20, we list the most common outputs observed. Each type of response accounts for a significant proportion of the models' answers across different questions. We found that although the model's outputs varied slightly with different prompts, a significant portion of the responses remained completely identical.

Prompt for Uniqueness Test and Sentence Grounding

Uniqueness Test
You are given two questions, the second question is based on the first question. You are also provided with the sentence containing the answer to the first question. Please identify which sentence contains the answer to the second question.

Question 1: `<Question 1>`
Answer to Question 1: `<Answer to Q1>`
Question 2: `<Question 2>`

Answer Candidates to Question 2:
0: `<Candidate 0>`
1: `<Candidate 1>`

Index Number of the Answer:

.....

Sentence Grounding
You are given a movie context and a question about it. Identify two sentences to answer the following question, which consists of two parts.

Note: First, indicate which sentence contains the answer to Part 1, and then specify which sentence contains the answer to Part 2. Your answer format should be: 'Sentence `<number>` for Part 1 and Sentence `<number>` for Part 2.'

Context: `<Context>`
Question: `<Question>`

Figure 17: The prompt for testing the uniqueness.

Task Instruction Prompts

1-Needle Questions
You are given a movie context consisting of several sentences and a series of consecutive scenes from the movie. Each scene is composed of five images stitched together from top to bottom. Identify which scene contains the necessary clues to answer the following question.
Note: Please provide the index number of the scene (e.g., 1, 2, or 3) that contains the necessary information.
Context: `<Context>`
Question: `<Question>`

. . . . . .

Visual Grounding 2-Needle Questions (Forward)

You are given a movie context and a sequence of consecutive movie scenes, each composed of 5 images stacked vertically. Identify two scenes to answer the following question, which consists of two parts.

Visual Grounding 2-Needle Questions (Reverse)
You are given a movie context and a sequence of reverse-order consecutive movie scenes, each composed of 5 images stacked vertically. Identify two scenes to answer the following question, which consists of two parts.
Note: First, indicate which scene contains the answer to Part 1, and then specify which scene contains the answer to Part 2. Your answer format should be: 'Scene `<number>` for Part 1 and Scene `<number>` for Part 2.'
Context: `<Context>`
The question has two parts:
Part 1: `<Part1>`
Part 2: `<Part2>`

. . . . . .

Image Description 2-Needle Questions

You are given a movie context consisting of several sentences and a series of consecutive scenes from the movie. Each scene is composed of five images stitched together from top to bottom.
You will be asked a two-part question. The first part of the question is designed to help you identify the scene that contains the necessary information to answer the second part. Using the correct scene, accurately answer Part 2 by selecting the most precise answer from the four given options.
Note: Please answer only the second part of the multiple-choice question.
Context: `<Context>`
Question: `<Question>`

Figure 18: The evaluation prompts for different types of questions.

Prompts and Outputs

Prompt 1
You are given a movie context and a sequence of consecutive movie scenes, each composed of 5 images stacked vertically. Identify two scenes to answer the following question, which consists of two parts.
Note: First, indicate which scene contains the answer to Part 1, and then specify which scene contains the answer to Part 2. Your answer format should be: 'Scene <number> for Part 1 and Scene <number> for Part 2.'

Context: `<CONTEXT>`
The question has two parts:
Part 1: `<PART1>`
Part 2: `<PART2>`

Output Samples: Question: What advantage did Batman have when he initiated the brutal fight against Superman, and how was he able to gain such an advantage?
GT: Scene 7, Scene 3
Answer: Scene 1 for Part 1 and Scene 2 for Part 2.

Question: What event led to a memorial service being held in Metropolis, and how did such an event occur?
GT: Scene 5, Scene 2
Answer: Scene 1 for Part 1 and Scene 2 for Part 2.

. . . . . .

Prompt 2
You are given a movie context and a sequence of consecutive movie scenes, each composed of 5 images stacked vertically. Identify two scenes to answer the following question, which consists of two parts.
Note: First, indicate which scene contains the answer to Part 1, and then specify which scene contains the answer to Part 2. Your answer format should be: 'Scene 7 for Part 1 and Scene 2 for Part 2.'

Output Samples: Question: What advantage did Batman have when he initiated the brutal fight against Superman, and how was he able to gain such an advantage?
GT: Scene 7, Scene 3
Answer: Scene 7 for Part 1 and Scene 2 for Part 2.

Question: What event led to a memorial service being held in Metropolis, and how did such an event occur?
GT: Scene 5, Scene 2
Answer: Scene 7 for Part 1 and Scene 2 for Part 2.

Figure 19: Various test prompts, but the model fails to generate reasonable answers.

Prompts and Outputs

Prompt 3
You are given a movie context and a sequence of consecutive images, numbered starting from 1. Your task is to identify two images to answer the following questions.
Note: First, indicate which image contains the answer to Part 1, and then specify which image contains the answer to Part 2. Your answer format should be: 'Image No.<number> for Q1 and Image No.<number> for Q2.'

Output Samples: Q: What advantage did Batman have when he initiated the brutal fight against Superman, and how was he able to gain such an advantage?
GT: Scene 7, Scene 3
Answer: Image No. 1 for Q1 and Image No. 2 for Q2.

. . . . . .

Prompt 4
You are given a movie context and a sequence of consecutive images. Your task is to identify the image that best matches each question.
Note: In your answer to each question, explain why you chose this image and specify which image it is in the sequence.

Output Sample (excerpt): Q1: The image that best matches this question is Image No. 1. This image shows Batman holding a Kryptonite weapon, which is a significant advantage for him in the fight against Superman.
Q2: The image that best matches this question is Image No. 2. This image shows Batman wearing an armored exoskeleton, which allows him to withstand Superman's attacks.

. . . . . .

Prompt 5
You are given a movie context and a sequence of consecutive images. Your task is to find the answers to the following two questions by identifying the image that best matches each question.
Note: In your answer, first explain why you chose that image, then count the number of this image in the sequence (e.g., "This image is the 1st image in the sequence.").

Output Sample (excerpt): Q1: The image that best matches this question is the 1st one in the sequence. This image shows Batman holding a Kryptonite weapon, which is a significant advantage for him in the fight against Superman.
Q2: The image that best matches this question is the 2nd one in the sequence. This image shows Batman wearing an armored exoskeleton, allowing him to withstand Superman's attacks and gain the upper hand.

Figure 20: Various test prompts, but the model fails to generate reasonable answers (continued).

G Examples

We present some examples from the dataset in the Fig. 21 and Fig. 22, covering all types of questions.

	Cause	Effect	Cause	Effect
				
Cause Sentence	Feeling responsible, Galahad gives a medal for valor to Lee's widow and his young son, Eggsy, instructing them to call the number on the back if they ever need help.		Arthur pours poisoned brandy for Eggsy, intending to activate the poison later; however, Eggsy surreptitiously switches the glasses.	
Effect Sentence	When Eggsy is arrested for stealing a car, he uses the number on the medal to call for assistance.		When Eggsy refuses to join Arthur's side, Arthur activates the poison in the brandy, unknowingly drinking from the poisoned glass himself and dying.	
Noncausal 1-N Question	1. What does Harry Hart deliver to Lee's widow, Michelle, and his son, Eggsy? 2. Who calls the number on the medal after being arrested?		1. What does Arthur pour for Eggsy that contains a poison? 2. Who activates the poison when Eggsy refuses to join his side?	
Causal 1-N Question	Why did Eggsy have a phone number to call when he was arrested for stealing a car?		What action did Eggsy take that led to Arthur unwittingly killing himself?	
Bridge Entity	Number on the medal		Poisoned drink	
VG 2-N Question	What approach did Eggsy take to seek help when he was arrested for stealing a car, and why was he able to adopt such an approach?		What deadly plan led to Arthur's own death when Eggsy refused to join his side, and how did Eggsy thwart this deadly plan?	
ID 2-N Question	What approach did Eggsy take to seek help when he was arrested for stealing a car, and when the event that enables Eggsy to use this approach occurs, what pattern is on the young boy's sweater? A. Fair Isle/Nordic pattern B. Houndstooth C. Cable Knit D. Birdseye		What deadly plan led to Arthur's own death when Eggsy refused to join his side? When the deadly plan backfires, what color is the wall behind the portraits hanging in the room? A. Green B. Black C. Brown D. Gold	

Figure 21: The examples from our dataset.

	Cause	Effect	Cause	Effect
				
Cause Sentence	Having stolen Kryptonite from LexCorp, Batman prepares for a confrontation with Superman, crafting an armored exoskeleton and Kryptonite weaponry.		Because Gordon was unable to reach her in time, Rachel dies.	
Effect Sentence	Despite Superman's attempts at diplomacy, Batman, armed with Kryptonite, initiates and wins a brutal fight against him.		Following Rachel's death, the Joker finds Dent in a hospital and manipulates him into seeking revenge.	
Noncausal 1-N Question	1. What does Batman steal from LexCorp? 2. Who ultimately wins the fight between Superman and Batman?		1. Who was unable to save Rachel? 2. Who does the Joker find in a hospital?	
Causal 1-N Question	What did Batman do beforehand that enabled him to win his brutal fight against Superman?		What prevented Rachel from being saved, enabling the Joker to manipulate Dent into seeking revenge for her death?	
Bridge Entity	Kryptonite		Rachel's death	
VG 2-N Question	What advantage did Batman have when he initiated the brutal fight against Superman, and how was he able to gain such an advantage?		What tragic event led the Joker to manipulate Dent into seeking revenge, and who failed to prevent such a tragic event?	
ID 2-N Question	What advantage did Batman have when he initiated the brutal fight against Superman? When the event that leads to Batman having the advantage occurs, what distinct shape is visible on the kryptonite spear after it pierces the containment unit? A. A droplet B. A batarang C. A handprint D. A Kryptonian glyph		What tragic event led the Joker to manipulate Dent into seeking revenge? When the tragic event that leads to the man in the hospital seeking revenge occurs, what is the man in glasses wearing around his neck? A. A scarf B. A tie C. A necklace D. A badge	

Figure 22: The examples from our dataset (continued) .