SCVBench: A Benchmark with Multi-turn Dialogues for Story-Centric Video Understanding

Sisi You[1,3], **Bowen Yuan**[1] and **Bing-Kun Bao**[1,2✉]
[1]Nanjing University of Posts and Telecommunications
[2]Pengcheng Laboratory
[3]State Key Laboratory of Tibetan Intelligence
{ssyou, bingkunbao}@njupt.edu.cn, yuanbw0925@gmail.com

Abstract

Video understanding seeks to enable machines to interpret visual content across three levels: action, event, and story. Existing models are limited in their ability to perform high-level long-term story understanding, due to (1) the over-simplified treatment of temporal information and (2) the training bias introduced by action/event-centric datasets. To address this, we introduce SCVBench, a novel benchmark for story-centric video understanding. SCVBench evaluates LVLMs through an event ordering task decomposed into sub-questions leading to a final question, quantitatively measuring historical dialogue exploration. We collected 1,253 final questions and 6,027 sub-question pairs from 925 videos, constructing continuous multi-turn dialogues. Experimental results show that while closed-source GPT-4o outperforms other models, most open-source LVLMs struggle with story-centric video understanding. Additionally, our StoryCoT model significantly surpasses open-source LVLMs on SCVBench. SCVBench aims to advance research by comprehensively analyzing LVLMs' temporal reasoning and comprehension capabilities. Code can be accessed at https://github.com/yuanrr/SCVBench.

1 Introduction

Video understanding enables machines with human-like visual interpretation capabilities, supporting applications such as intelligent surveillance and autonomous driving, achieving human-level perception and response to visual information.

Video understanding can be stratified into three levels based on the abstraction of its content: action-level, event-level, and story-level, as shown in Figure 1 (a-c). Specifically, action-level understanding focuses on recognizing the specific behaviors of individual subjects, event-level understanding aims to parse sequences of actions that form higher-level activities, and story-level understanding seeks to capture and interpret the logical and causal relationships among a series of connected events, thus fully comprehending the described narrative within the video. Existing methods [Feichtenhofer *et al.*, 2019; Dwibedi *et al.*, 2019] utilize convolutional neural networks and 3D CNNs to extract features from consecutive frames, thereby enabling the identification of specific actions, and others [Liu *et al.*, 2022; Wang *et al.*, 2023a; Tong *et al.*, 2022] apply more sophisticated structures such as recurrent neural networks, long short-term memory networks, and transformers for event-level understanding. Specifically, Large-scale Vision-Language Models (LVLMs) integrate Large Language Models (LLMs) [Touvron *et al.*, 2023; Achiam *et al.*, 2023; Bai *et al.*, 2023] with visual processors [Radford *et al.*, 2021; Dosovitskiy, 2020; Dosovitskiy, 2020] to demonstrate significant advancements in event understanding, *e.g.,* LLAVA-Video [Zhang *et al.*, 2024b] utilizes Qwen2 [Bai *et al.*, 2023] for semantic understanding and converts each frame into hundreds of visual tokens to extract visual features. Furthermore, Chain-of-Thought (CoT)[Wei *et al.*, 2022] is leveraged to imitate human cognitive patterns, incrementally enhancing their understanding of events through a step-by-step reasoning process. However, existing approaches face challenges in addressing story-level understanding, which requires long-term reasoning and high-level analysis. Specifically, current methods often simplify temporal information and overlook long-term dependencies, limiting their ability to understand causal relationships between events and construct a coherent story. Nonetheless, to the best of our knowledge, story-level video understanding remains underexplored in the existing literature. Establishing a comprehensive evaluation benchmark tailored to assess the progress in story-level video understanding achieved by LVLMs is imperative.

Existing video benchmarks mainly focus on action and event understanding through single-turn text inputs, which limits their effectiveness in assessing the complex, temporally sequential aspects of video understanding. For instance, current Video QA datasets, such as MLVU[Zhou *et al.*, 2024], Next QA[Xiao *et al.*, 2021], TemporalBench[Cai *et al.*, 2024], and VideoMME[Fu *et al.*, 2024], predominantly query the temporal dynamics of actions within videos by questions like "What did this person do after picking up a cup?" Existing methods can only comprehend simple actions and individual events, struggling with the complex narratives that constitute story-level understanding. Thus, these datasets are valuable for short-term temporal reasoning, and fail to capture multi-step complex events, limiting deeper video con-

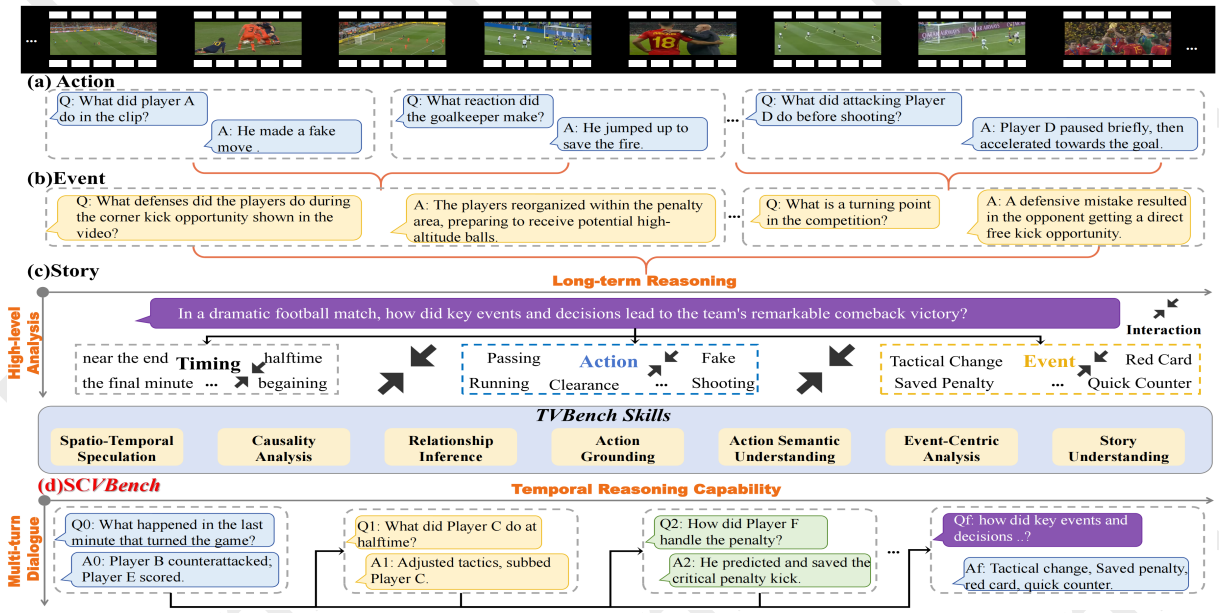

Figure 1: **Illustration of different-level video understanding.** The action-level analysis identifies individual behaviors like running or shooting. Event-level analysis parses sequences of actions, such as key moments in sports. Story-level interpretation captures the logical and causal relationships between connected events, revealing the overall narrative in the video content.

tent insights. Therefore, appropriate benchmarks are required to evaluate the ability to understand story-centric videos. To tackle these problems, we propose the first benchmark for story-centric video understanding, named **S**tory-**C**entric **V**ideo understanding **Bench**mark (**SCVBench**), which aims at comprehensively evaluating the high-level long-term understanding capabilities of LVLMs.

First, we propose a unique event ordering task to evaluate the story-centric video understanding capabilities of existing LVLMs, as shown in Figure 1 (d). This task requires models to sequence events based on their temporal relationships, assessing both action recognition and logical event connections. We innovatively decompose high-level story analysis into multi-turn dialogues, with each turn formulated as a focused sub-question that addresses specific segments or event characteristics. Specifically, the design of sub-questions considers diversity by covering various types of temporal logic, and coherence by forming a connected narrative, promoting deeper temporal understanding and reasoning abilities in the model. Furthermore, we can evaluate the model's ability to understand conversational context through whether utilize multi-turn dialogues for answering questions. Consequently, the designed task assesses the story-centric comprehension capabilities of LVLMs.

Second, we construct a large-scale dataset with temporal multi-turn question-answering dialogues for the proposed story-centric video understanding task. We compare our dataset with existing video datasets in Table 1. Our dataset comprises 925 diverse videos from YouTube, each undergoing thorough filtering and meticulous selection. The SCVBench dataset excels by incorporating multimodal information, maintaining contextual connections between QA pairs, and focusing questions on high-level story comprehension, thereby enabling a more comprehensive evaluation of models' integrated understanding and reasoning capabilities.

Third, we evaluate the video event ordering performance for various prevalent LVLMs on SCVBench. Our results provide the first comprehensive insight into the story-centric video understanding capabilities of existing LVLMs, surprisingly, these LVLMs fall far short of expectations. This motivates us to develop a stronger training-free LVLM, namely **StoryCoT**, which leverages an event extraction agent and story-centric reasoning agent to capture the nuances of events and their logical connections. Furthermore, extensive experimental results demonstrate that StoryCoT can significantly enhance the performance of LLAVA-OneVision [Li *et al.*, 2024a], Qwen2-VL [Bai *et al.*, 2023], and LLAVA-Video [Zhang *et al.*, 2024b].

2 Related Work

2.1 Large Vision-Language Models for video

Recently, Large Vision-Language Models (LVMs) have made significant strides in video understanding. The latest generative pre-trained transformer models feature advanced architectures and training methods, enhancing spatiotemporal feature capture in videos and improving semantic understanding through sophisticated multimodal fusion, *e.g.,* closed-source LVMs such as GPT-4V[Yang *et al.*, 2023], GPT-4o[Achiam *et al.*, 2023], and Gemini 1.5 Pro[Team *et al.*, 2024], and open-source alternatives like Video-ChatGPT[Maaz *et al.*, 2023], and VideoLLaMA2[Cheng *et al.*, 2024]. Models such as VideoMAE v2 [Wang *et al.*, 2023b] have further advanced the field by leveraging complex masking prediction strategies and temporal contrastive learning, thereby improving the utilization of unlabeled data and deepening the understand-

Dataset	#QAs	Avg. Q/V	Avg Dur.(s)	Mutlimodal	Dialogue	Story-centric	Annotation
EgoSchema [Mangalam *et al.*, 2023]	5,031	1.00	180	✓	✗	✗	Auto&Manual
ActivityNet-QA [Yu *et al.*, 2019]	800	1.00	180	✗	✗	✗	Manual
MVBench [Li *et al.*, 2024b]	4,000	1.10	16	✗	✗	✗	Auto
NExT-QA [Xiao *et al.*, 2021]	52,044	9.57	44	✗	✗	✗	Manual
LV-Bench [Wang *et al.*, 2024]	1,549	15.04	4,101	✗	✗	✗	Manual
LongVideoBench[Wu *et al.*, 2024]	6,678	1.77	473	✗	✓	✗	Manual
MLVU[Zhou *et al.*, 2024]	3120	3.35	720	✗	✗	✗	Manual
VideoMME[Fu *et al.*, 2024]	2,700	3.00	1,018	✓	✗	✗	Manual
SVBench[Yang *et al.*, 2025]	49,979	36.49	720	✓	✓	✗	Auto&Manual
SCVBench (Ours)	7,280	6.16	800	✓	✓	✓	Auto&Manual

Table 1: The comparison of different datasets. **Avg. Q/V**: the average number of QA pairs per video. **Avg Dur.(s)**: the average video length. **Multimodal**: whether the video consists of different modalities. **Dialogue**: whether there are contextual connections between QA pairs. **Story-centric**: whether the question focuses on the high-level story understanding.

ing of dynamic scenes. Furthermore, CoT-based video understanding decomposes complex video analysis tasks into a series of inference steps, constructing a step-by-step reasoning chain to guide the model toward final comprehension. Recent studies[Xu *et al.*, 2024; Dong *et al.*, 2024; Ni *et al.*, 2024] show that CoT prompts significantly enhance LLMs' reasoning and interpretability in video understanding, enabling more accurate capture and analysis of complex narrative structures within videos.

However, these LVLMs are still not fully adept at handling story-level video understanding and often fail to capture the complexities of real-world contexts. Moreover, current evaluation metrics primarily focus on sentence-level matching or overall segment similarity, which do not adequately measure the depth of complex story sequence comprehension. To rigorously evaluate the story understanding capabilities of these models, we propose SCVBench, a new benchmark designed to assess the performance of LVLMs in video-related tasks that imitate the complexity of real-world interactions.

2.2 Video Understanding Benchmarks

Existing question-answer benchmarks can be broadly categorized into two types to evaluate and advance video understanding technologies: action-event and event-centric datasets. The first category comprises traditional QA datasets such as TGIF-QA[Jang *et al.*, 2017], Next-QA[Xiao *et al.*, 2021], and MVBench[Li *et al.*, 2024b]. These datasets primarily focus on static or short video clips for question-answering tasks to test the model's understanding. For instance, TGIF-QA specializes in GIF-based QA, providing a large collection of animated GIFs paired with natural language questions that cover descriptive, transitional, and attribute-related queries. The second category includes datasets like VideoMME[Fu *et al.*, 2024], SVBench[Yang *et al.*, 2025],TemporalBench[Cai *et al.*, 2024], VidHal[Choong *et al.*, 2024], HourVideo[Chandrasegaran *et al.*, 2024], and MLVU[Zhou *et al.*, 2024], emphasizing action recognition and individual event. These datasets emphasize the temporal dynamics in videos through action/event-centric sequencing tasks, challenging models to accurately reconstruct or forecast event timelines. For example, HourVideo offers hour-long real-world video segments, featuring complex scenes with extended action sequences, enabling models to perform temporal ordering tasks in intricate environments.

Despite significant progress in these categories, these datasets are inadequate for evaluating models' ability to understand story-centric long videos. Specifically, traditional QA datasets focus on static or short clips, missing the complexity of evolving events, and temporal-centric datasets struggle to capture the causal relationships between various events. Therefore, we propose a novel and comprehensive benchmark SCVBench, specifically designed to evaluate the story-centric long video understanding abilities of models.

3 SCVBench

In this section, we propose a semi-automated annotation pipeline for videos, as shown in Figure 2, including a multi-stage LLM-assisted generation and curation process with several rounds of manual annotation.

3.1 Data Collection

To construct a dataset for evaluating the story-centric understanding capabilities of Large-scale Vision-Language Models (LVLMs), we selected publicly accessible amateur films of diverse genres from YouTube. These films offer two key advantages: (1) story complexity and character interaction, providing rich storylines and dense interactions that enhance the depth and accuracy of model evaluation; (2) information security assurance, publicly accessible amateur videos can inherently minimize the risk of personal information leakage and ensure data privacy.

We implemented rigorous collection and preprocessing steps to ensure data integrity and accuracy. We utilize yt-dlp to download videos, subtitles, and metadata, and WhisperX [Bain *et al.*, 2023] to fill in missing transcripts. Each video includes detailed metadata such as a synopsis providing a brief narrative overview, title, plot summary (logline), genre, release year, region, and language. To maintain the clarity and objectivity of the dataset, we retained only the synopsis that provides logically coherent story development, excluding potentially subjective or ambiguous descriptive elements such as the director's inspirations and actor perfor-

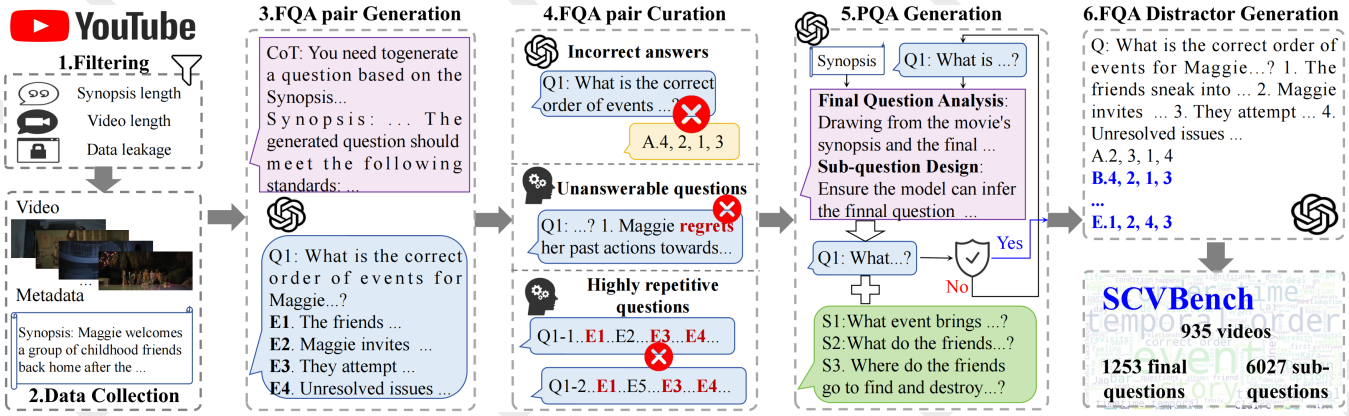

Figure 2: **Overview of the proposed SCVBench framework:** (1) Filtering raw videos from Youtube; (2) Data collection with ytdlp; (3) Generating final question-answer pair for each video; (4) Performing manual annotation and quality assessment for FQA pair curation; (5) Constructing progressive question-answering pairs; (6) Generating the distractor answers for FQA.

mances. The retained synopsis enables us to capture high-level, logically coherent event sequences while effectively avoiding the limitations of superficial action captions generated by models like Qwen2-VL [Bai *et al.*, 2023]. Additionally, this approach ensures that no unknown risks from LLM-generated content are introduced. Thus, by enhancing the authenticity and reliability of the dataset, we have established a robust foundation for evaluating the model's understanding capabilities and dialogue quality.

3.2 Question and Answer Generation

To foster advancements in story-centered video understanding, we have extended SCVBench with event ordering task under two complementary settings: Final Question-Answering (FQA) and Progressive Question-Answering (PQA), where PQA uses open-ended question answering to assist in selecting the correct answer for FQA. To populate these tasks, we first generate FQA pairs by guiding GPT-4 [Achiam *et al.*, 2023] with inputs including movie titles, loglines, and synopses. Subsequently, we perform auto and manual curation to refine the questions and answers, enhancing their relevance and accuracy. Moreover, a chain-of-thought prompting approach is utilized to facilitate the iterative generation and validation of sub-questions by GPT-4, ensuring that the PQA task effectively supports and enhances the FQA. The designed prompts in this part are detailed in supplement materials.

FQA pair Generation: For each film, we acquire human-authored metadata, including the movie title, logline, and synopsis, sourced directly from the corresponding YouTube channel. We then employ tailored prompts to instruct GPT-4 to generate pertinent question-answer pairs derived from these metadata. Specifically, we implement a 1-shot generation approach, emphasizing the exclusive use of provided synopsis to construct questions and answers, thus preventing any speculative assumptions. For each video, we generate 5 final questions. Each question is composed of 3 to 4 events randomly selected from the video's total of 7 to 8 events. This meticulous process yielded a dataset of 4,625 QA pairs, which form the basis of our multiple-choice question-answering task.

FQA pair Curation: To elevate the quality of the generated FQA pairs, we performed an in-depth analysis and identified three primary categories of errors: (1) the generated incorrect answers or formatting inaccuracies; (2) unanswerable questions based on synopsis, such as ambiguous or subjective ones; and (3) a high incidence of event repetition across 5 final questions derived from the same video. To tackle the first issue, we utilize GPT-4 to evaluate the generated questions against the provided synopsis and analyze the accuracy of the generated answers. This process identified and eliminated 133 final questions with incorrect answers. We adopt a manual screening approach to address the second and third issues. Specifically, we designed a Streamlit-based dataset management tool that enables users to perform database-like operations, such as deleting and modifying data entries. This tool enhances the review process by enabling efficient identification and removal of problematic questions, thereby improving the overall quality and coherence of the dataset. Finally, the manual screening process filtered 1,487 QA pairs.

PQA Generation: The Progressive Question-Answering (PQA) task aims to generate sub-questions that assist in answering the FQA task. It evaluates the model's ability to leverage historical dialogue information across multiple conversational turns by comparing performance with and without these sub-questions. Specifically, we employ a chain-of-thought method to guide GPT-4 in generating sub-questions based on the final question and the synopsis. The generation and validation of PQA constitute an iterative loop, in which sub-questions are systematically generated and rigorously validated. Samples that do not meet the validation criteria are re-generated, and this process iterates until all samples successfully pass validation.

The sub-questions generation is a two-stage process encompassing final question analysis and sub-question design. Firstly, we conduct an in-depth analysis of the movie synopsis to deduce the correct answer to the final question. Secondly, we design a set of sub-questions that collectively provide suf-

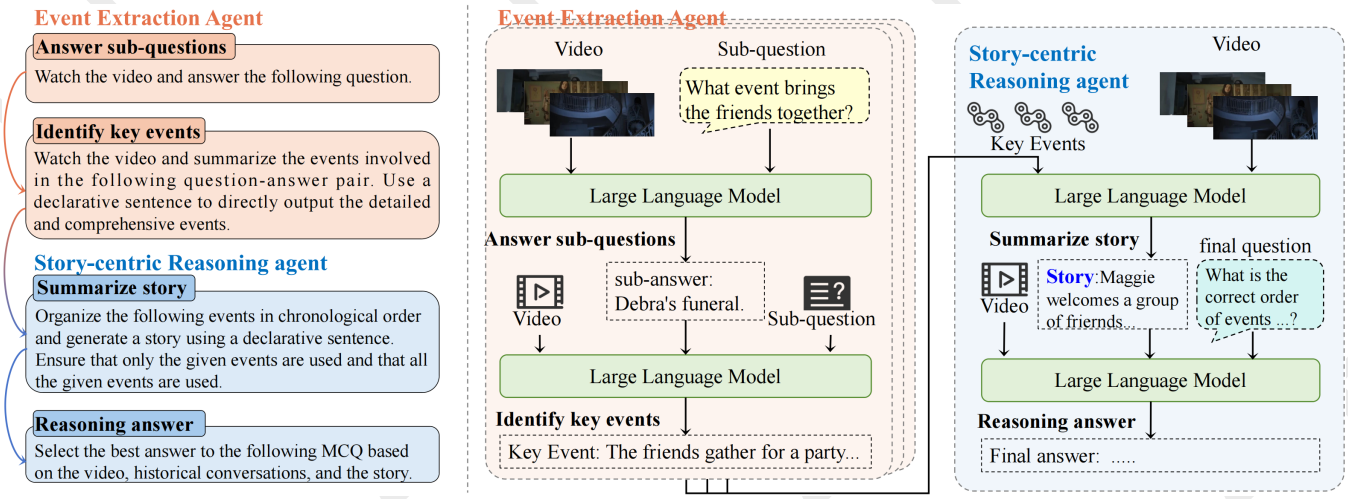

Figure 3: Architecture of the proposed StoryCoT model.

ficient information for inferring the final answer without requiring additional context from the movie. To ensure the sub-questions can effectively facilitate answering the final question, their generation is governed by several key principles:

- **Contextual Anchoring**: Each sub-question is anchored in the actions or environmental changes explicitly detailed within the movie synopsis.
- **Reasoning Facilitation**: The designed sub-questions facilitate deductive reasoning processes, maintaining a direct and meaningful association with the final question.
- **Selective Focus**: Particular attention is given to significant changes in the environment, actions undertaken, and behavioral modifications of characters, ensuring alignment with the overarching narrative.
- **Autonomous Structure**: Sub-questions are designed to be self-contained, eliminating interdependencies and sequential reliance. We further randomize sub-questions to prevent the model from relying on heuristic shortcuts based on sub-question order.
- **Temporal and Contextual Precision**: Sub-questions are crafted with explicit temporal markers and contextual boundaries to enhance clarity and specificity.
- **Conciseness and Specificity**: The generated answers are constrained to concise descriptions, typically limited to 3-5 words, promoting efficiency in both query formulation and response evaluation.

The PQA validation process assesses the model's performance in the initial chain-of-thought analysis. Samples that fail validation are cycled through additional generation-validation rounds, while those that succeed are directly added to the dataset. After 11 rounds, 33 unresolved samples due to complex or ambiguous final questions were removed.

FQA Distractor Generation: We design two experimental configurations to construct the Multiple-Choice Question (MCQ) task: the first involves questions that include five events, while the second concerns MCQs where the correct sequence of events must be selected. For the former configuration, we create distractors by randomly permuting numbers to construct plausible yet incorrect answer sequences. For the latter, a more sophisticated approach is adopted. Specifically, we utilize GPT-4 to generate four distractors for each question, based on the video synopsis, correct answer, and original question. The distractors are designed to appear plausible yet incorrect, maintaining syntactic similarity to the correct answer while introducing semantic distinctions. Simultaneously, we construct distractors by incorporating misdirection techniques such as character confusion or subtle plot adjustments, and introducing accurate but irrelevant details from other parts of the synopsis. The final dataset comprises 1,253 MCQs, each question consisting of one correct answer and four carefully constructed distractors.

3.3 Dataset Analysis and Statistics

The proposed SCVBench dataset comprises 925 distinct short films, with durations ranging from 5 to 37 minutes and an average length of 13 minutes. The cumulative runtime totals 240 hours, covering a variety of genres including rama (34.7%), science fiction (22.8%), horror (20.6%), comedy (12.1%), romance (3.0%), documentary (2.7%), animation (2.3%), and others (1.9%). Each film is accompanied by a title, a 15-word logline, and a 97-word synopsis, approximately two sentences in length.

For the video understanding evaluation, SCVBench includes 1,253 multiple-choice final questions and 6,027 open-ended sub-questions, averaging 1.35 final questions and 4.81 sub-questions per film. The question-answer pairs are diverse, with a median answer length of 5 words, and the dataset's vocabulary encompasses 8,928 unique terms.

Additionally, SCVBench features story-related questions that focus on the key sequence of events and character interactions within each synopsis and narrative arc. These questions assess participants' ability to understand logical coherence and interpret complex narratives, thereby evaluating the temporal video understanding capabilities of LVLMs.

Methods	Release Date	Visual Encoder	LLM	#Params	Frames	FQA Acc	PQA Acc	HDU
Random	-	-	-	-	-	20	20	-
Human Performance	-	-	-	-	-	93.3	93.3	-
Proprietary Models								
GPT-4V	Dec. 2023	-	GPT-4V	-	16	36.8	47.4	10.6
GPT-4o	May. 2024	-	GPT-4o	-	128	65.6	75.1	9.5
Gemini-1.5-Pro	Feb. 2024	-	Gemini-1.5-Pro	-	1fps	74.4	85.7	10.3
Open-Sourced Models								
LLAVA-OneVision	Aug. 2024	SigLIP/SO400M	Qwen2	0.5B	32	24.9	26.5	1.6
Qwen2-VL	Sep. 2024	ViT675M	Qwen2	2B	128	38.3	38.6	0.3
Oryx	Sep. 2024	OryxViT	Qwen2	7B	64	20.9	21.2	0.3
Oryx-1.5	Sep. 2024	OryxViT	Qwen2.5	7B	64	23.9	25.4	1.5
LLAVA-OneVision	Aug. 2024	SigLIP/SO400M	Qwen2	7B	32	42.6	51.2	8.6
LLAVA-Video	Oct. 2024	SigLIP/SO400M	Qwen2	7B	64	42.7	55.0	12.3
Qwen2-VL	Sep. 2024	ViT675M	Qwen2	7B	128	46.6	57.2	10.6

Table 2: Benchmark performance of LVLMs on our SCVBench dataset.

4 StoryCoT

To enhance reasoning capability through complex questions, we have developed a training-free multi-stage chain-of-thought method called StoryCoT. Unlike traditional approaches that rely heavily on extensive training data, StoryCoT innovatively employs a Multi-Agent System (MAS) featuring two specially designed agents: the Event Extraction Agent and the Story-Centric Reasoning Agent, as shown in Figure 3. These agents are responsible for processing information at different levels, facilitating the analysis of sub-questions, and providing effective assistance in addressing the final question.

Event Extraction Agent specializes in parsing specific events or event streams within each sub-question. It employs multiple independent Chain-of-Thought (CoT) to identify and summarize detailed event information or inter-event relationships from relevant video content. The generated event information consists of direct answers and sequenced events, establishing a robust foundation for subsequent high-level reasoning. Specifically, the Event Extraction Agent operates with a two-tier inference mechanism: firstly, it generates answers based on the sub-question and associated video content, integrating these answers into a multi-turn dialogue context; subsequently, it analyzes the sub-question along with its generated answer to extract concrete events or temporally ordered event sequences.

Story-Centric Reasoning Agent performs higher-level logical inferences based on the multiple event streams provided by the Event Extraction Agent. It synthesizes the answers from all sub-questions to construct a coherent storyline, thereby enhancing the response to the final question through this integrated information. This agent also encompasses a dual-layer reasoning process: initially, it summarizes the story by inferring the logical sequence of events; subsequently, it enhances the response to the final question using the summarized story information as historical context. This approach demonstrates significant potential in solving complex problems, effectively enhancing overall reasoning capabilities even without specialized training.

5 Experiments

5.1 Experimental Settings

Models. We evaluated seven LVLMs from different model families, including four open-sourced models with different parameters: LLAVA-OneVision [Li *et al.*, 2024a], Qwen2-VL [Bai *et al.*, 2023], Oryx [Liu *et al.*, 2024], and LLAVA-Video [Zhang *et al.*, 2024b] , and three proprietary models: GPT-4V [Yang *et al.*, 2023], GPT-4o [Achiam *et al.*, 2023] and Gemini-1.5 Pro [Team *et al.*, 2024]. These models exemplify a broad spectrum of architectural designs and training paradigms. To provide a comprehensive performance benchmark, we incorporated a random model, which selects candidate options through a stochastic process, and human performance metrics to establish an upper limit for comparison.

Implementation Details. We deploy our benchmark on Lmms-Eval [Zhang *et al.*, 2024a], an evaluation tool for diverse LVLMs. We perform standardized evaluations on our benchmark using greedy decoding for all LVLMs. All models are implemented following their reported default settings except for the Qwen2-VL series. Specifically, we report the results of the Qwen2-VL series with 128 input frames. For **FQA**, we use the instruction "Select the best answer to the following multiple-choice question based on the video and the historical conversations", while for **PQA**, we use "Select the best answer to the following multiple-choice question based on the video and the given story, as well as the historical conversations." We employ the post-prompt "Respond with only the letter (A, B, C, D, or E) of the correct option" to collect option responses directly.

Evaluation Metrics. To comprehensively evaluate the story-centric understanding capabilities of existing models, we have designed a suite of metrics based on SCVBench:

- **Final Question-Answer Accuracy (FQA Acc)**: This

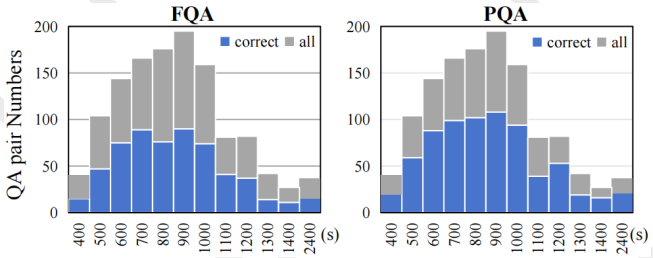

Figure 4: Performance of Qwen2-VL-7B on SCVBench under various video lengths.

metric measures the model's ability to directly and accurately answer the final question in a sequence.

- **Progressive Question-Answer Accuracy (PQA Acc)**: This metric assesses the model's performance in multi-turn dialogues, where it can infer the final answer from a series of sub-questions, reflecting its capacity for progressive reasoning over time.
- **Historical Dialogue Understanding (HDU)** aims to evaluate the capability of utilizing historical dialogue information, quantified by the difference between PQA Acc and FQA Acc. This metric evaluates how effectively the model leverages historical dialogue information to enhance its responses.

5.2 Overall Results

Benchmark Results. We present the overall performance of representative Large-scale Vision-Language Models (LVLMs) in Table 2, focusing on the event ordering task to evaluate their ability to understand story-centered videos. The following key observations can be drawn from this analysis: (1) Proprietary Model Superiority: Proprietary models like Gemini-1.5-Pro outperform open-source models, achieving accuracies of 74.4% for FQA Acc. (2) Model Scale Impact: Larger models (*e.g.*, 7B parameters) consistently show better performance than smaller ones, indicating that model size positively influences understanding and reasoning capabilities. (3) Multi-turn Dialogue Enhancement: Auxiliary information significantly improves the accuracy of each model, *e.g.,* LLAVA-Video's performance increasing from 42.7% to 55.0% with multi-turn dialogue. (4) Historical Dialogue Understanding: While multi-turn dialogues generally enhance model performance, the degree of improvement varies significantly. Models with larger parameter sizes and integrated visual encoders, such as LLAVA-Video, exhibit higher HDU metrics, indicating superior historical dialogue understanding. In contrast, smaller models like Oryx show limited improvements, primarily due to differences in architecture and training data design. (5) Visual Encoder Impact: Different visual encoders exhibit distinct reasoning capabilities,*e.g.,* LLAVA-Video to surpass Oryx by 21.8/33.8% on FQA/PQA, validating its superiority in long-video spatiotemporal modeling. (6) Human Performance Benchmark: Human accuracy sets a benchmark, showing that while some models perform well, there is still room for improvement, especially in complex scenarios. Furthermore, we illustrated the performance of Qwen2-VL-7B across various video lengths, as shown in Figure 4. As video length increases, the correct pairs tend to decline, and PQA task outperforms FQA across various video lengths, indicating that a step-by-step reasoning approach contributes to better story comprehension.

Methods	**PQA Acc**	**PQA Acc +*StoryCoT***	**PQA Acc +Sub-GT**
LLAVA-OneVision-0.5B	26.5	28.3	29.2
Qwen2-VL-2B	38.6	39.5	42.2
LLAVA-OneVision-7B	51.2	53.3	66.0
LLAVA-Video-7B	55.0	57.6	67.2
Qwen2-VL-7B	57.2	58.7	69.4

Table 3: Effect of StoryCoT.

In conclusion, existing models demonstrate considerable limitations in story-centric video understanding tasks, with most methods achieving FQA performance below 50%. This shortfall is primarily due to the task's complexity, requiring models to process dynamically evolving video content, continuously track complex visual elements, and integrate them into a coherent narrative. These demands challenge temporal sequence understanding, extended context handling, and continuous scene comprehension. Future research should focus on refining architectures, expanding datasets, and enhancing task complexity to improve performance and move closer to human-level comprehension.

Effect of StoryCT. We evaluated models on PQA tasks using two methods: PQA + StoryCOT, which enhances narrative understanding through story coherence cues, and PQA + Sub-GT, which uses annotated sub-questions to improve inference. As shown in Table 3, it can be reserved that StoryCOT significantly boosts PQA accuracy across all models, exemplified by LLAVA-OneVision-7B increasing from 51.2% to 53.3%. Additionally, PQA + Sub-GT demonstrates the effectiveness of sub-questions, with LLAVA-OneVision-7B's accuracy jumping from 51.2% to 66.0%. In summary, StoryCOT universally enhances PQA performance, while Sub-GT highlights the importance of sub-questions in improving model accuracy.

6 Conclusion

This paper presents SCVBench, a new benchmark for evaluating story-centered video understanding. SCVBench includes 925 videos from YouTube, 1,253 annotated final question-answer pairs, and 6,027 sub-questions to form multi-turn dialogues. Our experiments show that while current state-of-the-art LVLMs perform well in event-level video understanding, they do not reach human-level accuracy in understanding story-focused videos. To address these limitations, we developed the StoryCoT model, leveraging event extraction and story-centric reasoning agents to guide step-by-step reasoning and progressively reach a well-reasoned final answer. Our approach significantly outperforms existing open-source LVLMs on SCVBench, aiming to spur the development of advanced models capable of handling story-centered video complexities.

Acknowledgments

This work was supported by the National Natural Science Foundation of China under Grants 62301276 and 62325206, the Key Research and Development Program of Jiangsu Province under Grant BE2023016-4, and the Opening Foundation of the State Key Laboratory of Tibetan Intelligence, Key Laboratory of Tibetan Information Processing, Ministry of Education (2024-2-003).

References

[Achiam *et al.*, 2023] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[Bai *et al.*, 2023] Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. Qwen technical report. *arXiv preprint arXiv:2309.16609*, 2023.

[Bain *et al.*, 2023] Max Bain, Jaesung Huh, Tengda Han, and Andrew Zisserman. Whisperx: Time-accurate speech transcription of long-form audio. *arXiv preprint arXiv:2303.00747*, 2023.

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

[Chandrasegaran *et al.*, 2024] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Li Fei-Fei. Hourvideo: 1-hour video-language understanding. *arXiv preprint arXiv:2411.04998*, 2024.

[Cheng *et al.*, 2024] Zesen Cheng, Sicong Leng, Hang Zhang, Yifei Xin, Xin Li, Guanzheng Chen, Yongxin Zhu, Wenqi Zhang, Ziyang Luo, Deli Zhao, et al. Videollama 2: Advancing spatial-temporal modeling and audio understanding in video-llms. *arXiv preprint arXiv:2406.07476*, 2024.

[Choong *et al.*, 2024] Wey Yeh Choong, Yangyang Guo, and Mohan Kankanhalli. Vidhal: Benchmarking temporal hallucinations in vision llms. *arXiv preprint arXiv:2411.16771*, 2024.

[Dong *et al.*, 2024] Yuhao Dong, Zuyan Liu, Hai-Long Sun, Jingkang Yang, Winston Hu, Yongming Rao, and Ziwei Liu. Insight-v: Exploring long-chain visual reasoning with multimodal large language models. *arXiv preprint arXiv:2411.14432*, 2024.

[Dosovitskiy, 2020] Alexey Dosovitskiy. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[Dwibedi *et al.*, 2019] Debidatta Dwibedi, Yusuf Aytar, Jonathan Tompson, Pierre Sermanet, and Andrew Zisserman. Temporal cycle-consistency learning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1801–1810, 2019.

[Feichtenhofer *et al.*, 2019] Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211, 2019.

[Fu *et al.*, 2024] 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.

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

[Li *et al.*, 2024a] 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*, 2024.

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

[Liu *et al.*, 2022] Ze Liu, Jia Ning, Yue Cao, Yixuan Wei, Zheng Zhang, Stephen Lin, and Han Hu. Video swin transformer. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3202–3211, 2022.

[Liu *et al.*, 2024] Zuyan Liu, Yuhao Dong, Ziwei Liu, Winston Hu, Jiwen Lu, and Yongming Rao. Oryx mllm: On-demand spatial-temporal understanding at arbitrary resolution. *arXiv preprint arXiv:2409.12961*, 2024.

[Maaz *et al.*, 2023] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023.

[Mangalam *et al.*, 2023] 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.

[Ni *et al.*, 2024] Minheng Ni, Yutao Fan, Lei Zhang, and Wangmeng Zuo. Visual-o1: Understanding ambiguous instructions via multi-modal multi-turn chain-of-thoughts reasoning. *arXiv preprint arXiv:2410.03321*, 2024.

[Radford *et al.*, 2021] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack

Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021.

[Team *et al.*, 2024] 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.

[Tong *et al.*, 2022] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022.

[Touvron *et al.*, 2023] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Wang *et al.*, 2023a] Jinpeng Wang, Yixiao Ge, Rui Yan, Yuying Ge, Kevin Qinghong Lin, Satoshi Tsutsui, Xudong Lin, Guanyu Cai, Jianping Wu, Ying Shan, et al. All in one: Exploring unified video-language pre-training. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6598–6608, 2023.

[Wang *et al.*, 2023b] Limin Wang, Bingkun Huang, Zhiyu Zhao, Zhan Tong, Yinan He, Yi Wang, Yali Wang, and Yu Qiao. Videomae v2: Scaling video masked autoencoders with dual masking. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14549–14560, 2023.

[Wang *et al.*, 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*, 2024.

[Wei *et al.*, 2022] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

[Wu *et al.*, 2024] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *arXiv preprint arXiv:2407.15754*, 2024.

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

[Xu *et al.*, 2024] Guowei Xu, Peng Jin, Li Hao, Yibing Song, Lichao Sun, and Li Yuan. Llava-o1: Let vision language models reason step-by-step. *arXiv preprint arXiv:2411.10440*, 2024.

[Yang *et al.*, 2023] Zhengyuan Yang, Linjie Li, Kevin Lin, Jianfeng Wang, Chung-Ching Lin, Zicheng Liu, and Lijuan Wang. The dawn of lmms: Preliminary explorations with gpt-4v (ision). *arXiv preprint arXiv:2309.17421*, 9(1):1, 2023.

[Yang *et al.*, 2025] Zhenyu Yang, Yuhang Hu, Zemin Du, Dizhan Xue, Shengsheng Qian, Jiahong Wu, Fan Yang, Weiming Dong, and Changsheng Xu. Svbench: A benchmark with temporal multi-turn dialogues for streaming video understanding. *arXiv preprint arXiv:2502.10810*, 2025.

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

[Zhang *et al.*, 2024a] Kaichen Zhang, Bo Li, Peiyuan Zhang, Fanyi Pu, Joshua Adrian Cahyono, Kairui Hu, Shuai Liu, Yuanhan Zhang, Jingkang Yang, Chunyuan Li, and Ziwei Liu. Lmms-eval: Reality check on the evaluation of large multimodal models, 2024.

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

[Zhou *et al.*, 2024] 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.