Reasoning is All You Need for Video Generalization: A Counterfactual Benchmark with Sub-question Evaluation

Qiji Zhou[* 1], **YiFan Gong**[* 2], **Guangsheng Bao**[1], **Hongjie Qiu**[2],
Jinqiang Li[2] **Xiangrong Zhu**[2], **Huajian Zhang**[1], **Yue Zhang**[† 1]

[1]School of Engineering, Westlake University
[2]College of Computer Science and Technology, Hangzhou Dianzi University
{zhouqiji, baoguangsheng, zhanghuajian, zhangyue}@westlake.edu.cn
{gongyifan, qiuhongjie, lijinqiang, zhuxiangrong}@hdu.edu.cn

Abstract

Counterfactual reasoning is crucial for robust video understanding but remains underexplored in existing multimodal benchmarks. In this paper, we introduce **COVER** (**CO**unterfactual **V**id**E**o **R**easoning), a multidimensional multimodal benchmark that systematically evaluates MLLMs across the abstract-concrete and perception-cognition dimensions. Beyond prior multimodal benchmarks, COVER decomposes complex queries into structured sub-questions, enabling fine-grained reasoning analysis. Experiments on commercial and open-source models reveal a strong correlation between sub-question accuracy and counterfactual reasoning performance, highlighting the role of structured inference in video understanding. Furthermore, our results suggest a key insight: enhancing the reasoning capability of models is essential for improving the robustness of video understanding. COVER establishes a new standard for assessing MLLMs' logical reasoning abilities in dynamic environments. Our work is available at https://github.com/gongyifan-hash/COVER-Benchmark.

1 Introduction

In recent years, the rapid advancement of large language models (LLMs) has spurred growing interest in multimodal large language models (MLLMs) (Hurst et al., 2024; Anthropic, 2024; Chen et al., 2024; Zhang et al., 2024a, 2025; Wang et al., 2024; Wu et al., 2024b). Various early benchmarks have been proposed to assess multimodal understanding ability of MLLMs, particularly in static images (Fu et al., 2023; Hudson and Manning, 2019; Liu et al., 2024; Yu et al., 2024). More recently, benchmarks involving complex images and dynamic videos have emerged to evaluate MLLM's capabilities in temporal reasoning, spatio-temporal recognition, and object detection (Fu et al., 2024;

Li et al., 2024b, 2023). Despite these advances, these benchmarks often overlook counterfactual reasoning, which is a critical component for evaluating inference in complex and realistic environments. As a result, they fall short of providing a comprehensive assessment of MLLMs' reasoning capabilities.

Counterfactual reasoning, which posits hypothetical alternatives to observed realities, is pivotal for advanced video inference and is closely tied to out-of-distribution generalization (Yang et al., 2023; Bao et al., 2025). Previous work has attempted to construct counterfactual queries for images and videos (Li et al., 2024d,e,c; Patel et al., 2022; Wu et al., 2023). Most existing multimodal counterfactual benchmarks tend to focus on assessing subtask-specific robustness of reasoning ability (Li et al., 2024e; Wu et al., 2024c, 2023). However, they do not assess the underlying factors that contribute to the robustness of these reasoning capabilities. Such benchmarks often lack a systematic progression from abstract to concrete dimensions and from low-level perception to high-level cognition, limiting their ability to comprehensively capture multimodal reasoning processes in MLLMs. Furthermore, these benchmarks rarely investigate how robust video understanding interacts with stepwise reasoning in dynamic environments, leaving a gap in our assessment of advanced inference skills.

To bridge this gap, we propose **COVER**, a counterfactual video reasoning benchmark driven by a multidimensional abstraction level evaluation mechanism. Unlike existing multimodal counterfactual benchmarks, which often focus on multitask-oriented questions, COVER systematically classifies tasks into four quadrants. We define specific tasks for each quadrant to evaluate MLLMs' diverse reasoning capabilities in complex video scenarios. Beyond merely posing counterfactual questions, COVER introduces a *sub-question* reasoning mechanism derived from necessary con-

---

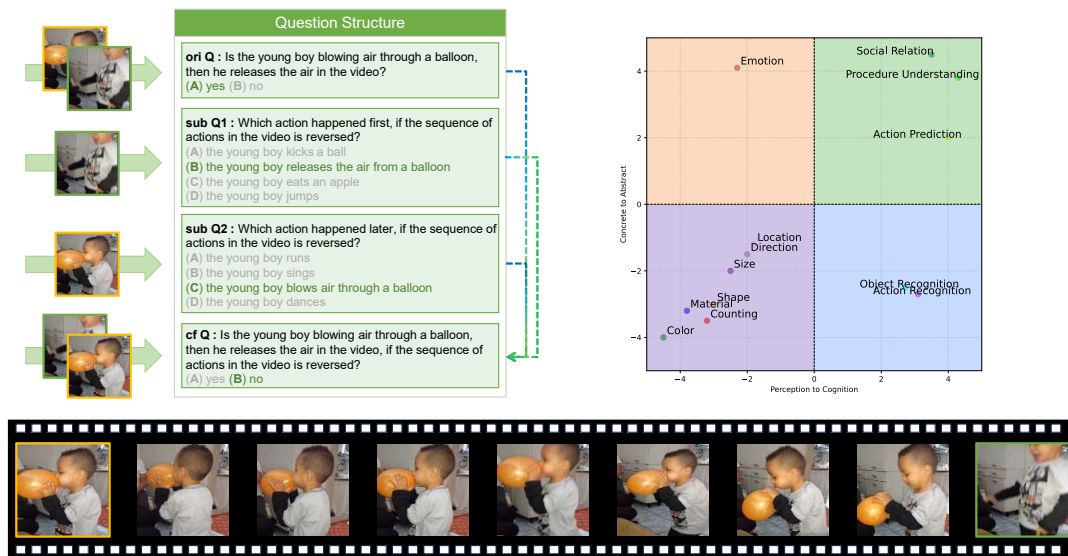[*]Equal contribution.
[†]Corresponding Author.

Figure 1: An example from the COVER benchmark. The ground-truth answers are highlighted in green. All data—including original questions, counterfactual questions, sub-questions, and videos—have been manually verified as part of COVER. The diagram in the upper right corner illustrates the division of each COVER task into four quadrants.

ditions, enabling a deeper evaluation of performance across MLLMs. This approach allows us to establish a connection between the accuracy of intermediate steps and the overall robustness of counterfactual reasoning. As shown in Figure 1, when asked to determine whether a boy completes a series of actions in a specified order, COVER decomposes the problem into multiple steps, each representing a necessary condition. For instance, sub-question *Q1* may inquire about which action occurs first in the reversed video, while sub-question *Q2* targets the final action. This structured approach not only helps evaluate how a model adapts to event-sequence changes but also reveals its strengths and weaknesses in extracting and synthesizing critical information under counterfactual assumptions. By encompassing a broad range of abstraction levels, COVER stands as the most comprehensive dataset of its kind, paving the way for more rigorous and holistic evaluations of MLLMs' dynamic and counterfactual reasoning capabilities.

Building on the COVER benchmark, we conducted a series of systematic experiments using both open-source and commercial closed-source models of varying scales. Our results indicate a strong positive correlation between the models' sub-question accuracy and performance in counterfactual reasoning and robust video understanding. The findings underscore the tight linkage

between sophisticated inference capabilities and high-level video comprehension. Furthermore, we examine how automatically generated versus human-guided sub-question decomposition (chain-of-thought, CoT (Wei et al., 2022)) influences complex reasoning and identifies the key factors impacting inference accuracy in MLLMs. Through these experiments, COVER offers valuable insights into how structured reasoning can enhance the robustness of video understanding by constructing a sub-question–based counterfactual video QA benchmark across multiple levels of abstraction and thoroughly evaluating mainstream MLLMs' logical reasoning abilities.

2 Related Work

Multimodal Large Language Models and Their Evaluation. Recent advances in MLLMs have greatly improved their capacity to understand and reason over diverse modalities, such as images, text, and videos. To evaluate these models, benchmarks targeting static image comprehension have emerged, including MM-Vet (Yu et al., 2024), MME (Fu et al., 2023), MMBench (Liu et al., 2024), and GQA (Hudson and Manning, 2019). These primarily assess visual recognition and spatial reasoning. Extending beyond static content, video-centric benchmarks like Video-MME (Fu et al., 2024), MvBench (Li et al., 2024b), and

Benchmark	Video	Q&A	Qs Source	CF	SQP	PCD	ACD
CoFCA (Wu et al., 2024a)	✗	✓	H&A	✓	✓	✗	✗
CFMM (Li et al., 2024e)	✗	✓	H	✓	✗	✗	✗
Video-MME (Fu et al., 2024)	✓	✓	H	✗	✗	✓	✗
CRIPP-VQA (Patel et al., 2022)	✓	✓	H	✓	✗	✗	✗
VITATECS (Li et al., 2024c)	✓	✗	H&A	✓	✗	✗	✗
COVER (ours)	✓	✓	H&A	✓	✓	✓	✓

Table 1: Comparison with existing benchmarks. **Video**: whether the benchmark involves video data; **Q&A**: whether it follows a question-and-answer format; **Qs source**: H indicates human annotation, A indicates automatic annotation; **CF**: whether counterfactual questions are included; **PCD**: whether the benchmark is categorized by the model's perceptual and cognitive demands; **ACD**: whether tasks are divided based on object abstraction (abstract vs. concrete).

SEED-Bench (Li et al., 2023) focus on temporal dynamics and contextual reasoning. Together, these benchmarks reflect the growing demand for evaluating multimodal understanding in both static and dynamic environments.

Chain-of-Thought and Counterfactual Reasoning in MLLMs. Chain-of-Thought (CoT) reasoning (Wei et al., 2022) enhances logical inference by breaking down complex tasks into intermediate steps. Multimodal adaptations (Zhang et al., 2024b; Zheng et al., 2023) extend this strategy across modalities, showing gains in structured reasoning. Counterfactual reasoning, which examines hypothetical changes and their consequences, has also gained traction. Prior work explores this in text (Wu et al., 2024c,a), visual QA (Li et al., 2024e), and hybrid settings. ACQUIRED (Wu et al., 2023) proposes a taxonomy of counterfactual types, while AuroraCap (Chai et al., 2024) and CoFCA (Wu et al., 2024a) assess models' sub-task decomposition and multi-step reasoning. These approaches collectively underscore the importance of structured, causal reasoning in complex multimodal tasks.

Multimodal Generalization and Video Counterfactual Benchmarks. Although several benchmarks target video-based counterfactual understanding—such as CRIPP-VQA for physical properties, VITATECS for captioning, and ACQUIRED for scenario taxonomy (Li et al., 2024c; Patel et al., 2022)—they remain narrow in scope. Most fail to capture the breadth of reasoning demands in real-world counterfactual scenarios.

To address this, COVER introduces a fine-grained framework for evaluating counterfactual video reasoning via sub-question decomposition. It explicitly distinguishes between abstract vs. concrete ob-

ject attributes and perceptual vs. cognitive reasoning demands. As summarized in Table 1, COVER broadens the evaluation spectrum, enabling a more nuanced and comprehensive assessment of multimodal counterfactual reasoning than prior efforts.

3 The COVER Benchmark

Figure 2: Overview of the 13 tasks in COVER. Numbers on the outer edge of the rose chart indicate the total number of question pairs for each task, while inner labels denote the corresponding dimension: **A&C** (Abstract Cognition), **C&C** (Concrete Cognition), **A&P** (Abstract Perception), and **C&P** (Concrete Perception).

This section provides a comprehensive overview of the construction of **COVER**. We introduce our data partitioning framework designed to evaluate MLLM reasoning ability across four complementary dimensions. Next, we describe the data curation process, which domain experts have rigorously validated to ensure the high quality and reliability of the benchmark.

Our benchmark includes approximately 2,800

videos, which are paired with around 12,000 to 13,000 individual QA instances. As stated in L-Figure 6, the enhanced version of our dataset consists of about 2.9k question pairs, with each pair comprising at least three individual QA items: one original question, one counterfactual question, and at least one sub-question (often multiple).

3.1 Benchmark Definition

As illustrated in Figure 2, we categorize the 13 benchmark tasks into four quadrants based on the abstract-concrete and perceptual-cognitive dimensions. **Abstract-Perception:** (1) Emotion: Understanding and recognizing emotional states. **Concrete-Perception:** (2) Counting: Quantity recognition and calculation. (3) Color: Perceiving object colors. (4) Direction: Sensing motion trends. (5) Size: Identifying object dimensions. (6) Shape: Perceiving object shapes. (7) Material: Recognizing object materials. (8) Location: Detecting object positions. **Concrete-Cognition:** (9) Action Recognition: Identifying specific actions. (10) Object Recognition: Recognizing specific objects. **Abstract-Cognition:** (11) Action Prediction: Forecasting future actions. (12) Procedure Understanding: Comprehending sequential processes and logic. (13) Social Relation: Understanding social relationships.

Division of Abstract and Concrete Scenes. This distinction reflects a functional differentiation within cognitive representation systems. Neuroscientific studies (Katja Wiemer-Hastings and Xu, 2005) suggest that concrete concepts rely heavily on multi-modal perceptual simulations (e.g., object shape, material), while abstract concepts are primarily represented through language-mediated symbolic operations. Abstract tasks often require integrating non-perceptual information, such as contextual encoding for emotion recognition or constructing temporal causal models for action prediction.

Division of Perception and Cognition. Perception involves the initial reception of external stimuli through sensory organs, converting them into neural signals that provide raw environmental data. Cognition, built upon perception, refers to the further processing of these signals, encompassing higher-level mental functions such as memory, attention, language comprehension, problem-solving, and reasoning. This distinction underscores different stages of information processing, with perception forming the foundation upon which cognitive

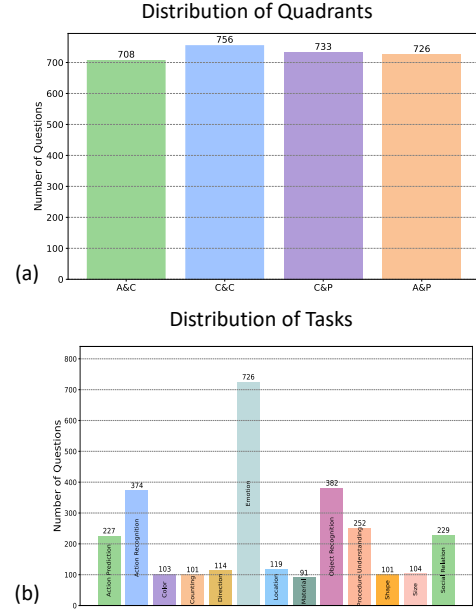functions are built.

3.2 Data Construction

Figure 3: (a) Distribution of question pairs across the four quadrants. (b) Distribution of question pairs across the 13 tasks.

To construct COVER, we carefully selected a diverse range of open-source and research-available video sources, including Sigurdsson et al. (2016); Yi et al. (2020); Xie et al. (2024); Tan et al. (2020); Shahroudy et al. (2016); Pătrăucean et al. (2023); Zhang et al. (2023); Gao et al. (2017); Jang et al. (2017); Wang et al. (2019); Krantz et al. (2020). These sources encompass various real-world scenarios, ranging from daily activity recognition to complex scene understanding. As shown in Appendix Figure 6, we collected 146 videos and designed 150 aspect-specific QA pairs, each of which underwent dual-team review for validation. To ensure balanced coverage across the four quadrants, we expanded the seed data using GPT-generated instances (720-760 per quadrant) to mitigate any potential biases. The detailed statistical findings are comprehensively presented in Figure 3. The frame count of videos in **COVER** ranges from 16 to 1739, with an average of 294.34 frames. We finally constructed 2,923 high-quality counterfactual question-answer pairs. Each question-answer pair consists of an original question, which presents no hypothetical context, and a counterfactual question, which introduces situational assumptions and sub-

	ori_{acc}	cf_{acc}	sub_{acc}
GPT-4o	70.26	45.93	56.94
GPT-4o-mini	67.32	51.47	55.94
Claude-3.5-Sonnet	63.60	38.04	49.40
Gemini-1.5-Pro	74.82	49.64	63.76
Gemini-1.5-Flash	73.90	48.75	62.52
Gemini-2.0-Flash	**77.18**	46.90	62.92
InternVL2.5-78B	76.74	**59.46**	**67.23**
LlaVA-Video-72B	64.35	56.04	61.54
InternVL2.5-26B	75.40	51.08	62.65
InternVL2.5-8B	74.31	57.75	61.65
VideoLlama3-8B	73.04	51.25	60.09
LlaVa-OV-7B	62.74	51.80	56.42
LLaVA-Video-7B	60.52	51.93	55.11
Qwen2-VL-7B	71.83	46.90	58.40
VILA-U-7B	60.01	38.42	47.32
VILA1.5-7B	60.25	57.34	53.18

Table 2: General assessment results of COVER. ori_{acc}, cf_{acc}, and sub_{acc} denote the accuracies of the original, counterfactual, and sub-questions, respectively.

questions that enable granular reasoning analysis.

Eight annotators further validated the dataset and checked logical consistency to ensure the reasoning relied solely on the video content. Additionally, three experts cross-validate the dataset (see Appendix Table 9) to confirm the structural balance.

4 Experiments

In this section, we systematically evaluate MLLMs of varying scales on the COVER benchmark to foster transparent and reproducible research. Our evaluation spans four key dimensions: cognition, perception, abstraction, and concreteness. It encompasses diverse reasoning sub-tasks, including counterfactual reasoning, direct inference, and sub-question-guided reasoning. We compare both open-source and proprietary models across different parameter scales to analyze their relative strengths and limitations. We begin by detailing the experimental setup.

4.1 Settings

To ensure a thorough evaluation, we selected a diverse set of representative MLLMs, including commercial closed-source models such as GPT-4o (Hurst et al., 2024), Claude (Anthropic, 2024), and Gemini (Reid et al., 2024), as well as leading open-source models such as InternVL2.5 (Chen et al., 2024), LLAVA-Video (Zhang et al., 2024a), LLaVA-OV (Li et al., 2024a), Qwen2-VL (Wang et al., 2024), VideoLLaMA3 (Zhang et al., 2025), vila-u (Wu et al., 2024b), and VILA (Lin et al., 2024). These models span a wide range of pa-

rameter scales and design paradigms, offering a comprehensive view of the current landscape in multimodal learning.

We evaluate model performance on video understanding using three metrics: ori_{acc} (original question accuracy), cf_{acc} (counterfactual question accuracy), and sub_{acc} (sub-question accuracy), with scores averaged over at least three runs. All models are tested under identical conditions, using a consistent frame extraction strategy that samples 16 frames per video segment. The impact of alternative sampling strategies is discussed in Chapter 5.

4.2 Main Results

As shown in Table 2, Gemini-2.0-Flash (ori_{acc} 77.18%) and InternVL2.5-78B (ori_{acc} 76.74%) rank as the top two models, demonstrating their strong foundational video understanding. The lower scores of VILA-U-7B (ori_{acc} 60.01%) and LLaVA-Video-7B (ori_{acc} 60.52%) highlight the limitations of smaller models in processing long sequences. InternVL2.5-78B (cf_{acc} 59.46%) shows significant dominance in handling conditional reasoning and complex contexts. Notably, counterfactual questions cause sharp accuracy drops compared to the original questions: GPT-4o (-24.33%) and Gemini-1.5-Pro (-25.18%), indicating that most models struggle with counterfactual reasoning.

Most models exhibit higher sub_{acc} than cf_{acc} (e.g., Claude-3.5-Sonnet 49.40% vs. 38.04%, LLaVA-Video-72B 61.54% vs. 56.04%). This suggests better stability in localized reasoning tasks than in holistic tasks, where error accumulation impacts performance. In the Appendix, we provide detailed case demonstrations in Figure 8.

Closed-source Model Performance. As shown in Table 3, Gemini 1.5 Pro demonstrates strong dominance in both concrete cognition (ori_{acc} 82.14%) and abstract perception tasks (ori_{acc} 75.48%). Gemini 2.0 Flash excels in abstract perception (ori_{acc} 75.90%) and concrete perception tasks (ori_{acc} 74.22%), showcasing strong capabilities in handling high-complexity perceptual tasks.

Open-source Model Performance. As shown in Table 3, InternVL2.5-78B leads in abstract cognition (ori_{acc} 72.88%) and concrete perception tasks (cf_{acc} 58.25%), reflecting a deep understanding of abstract concepts and complex logic. Lightweight models like Qwen2-VL-7B perform well in concrete cognition (ori_{acc} 82.14%) but face limitations in abstract tasks (ori_{acc} 65.96% in A&C) due to

Models	A&C (%)			C&C (%)			C&P (%)			A&P (%)		
	ori_{acc}	cf_{acc}	sub_{acc}	ori_{acc}	cf_{acc}	sub_{acc}	ori_{acc}	cf_{acc}	sub_{acc}	ori_{acc}	cf_{acc}	sub_{acc}
GPT-4o	71.05	41.81	41.70	74.87	43.65	68.36	69.95	42.62	50.52	65.01	55.65	63.97
GPT-4o-mini	62.29	52.40	42.97	76.32	54.37	66.49	64.62	40.85	44.78	65.56	**58.26**	65.96
Claude-3.5-sonnet	56.92	37.01	35.55	70.11	42.33	61.77	60.03	32.88	40.08	66.94	39.81	56.81
Gemini 1.5 Pro	69.49	44.49	<u>53.36</u>	<u>82.14</u>	51.98	72.78	71.76	43.93	56.81	75.48	57.99	69.54
Gemini 1.5 Flash	70.48	45.34	52.23	82.01	49.34	71.51	70.67	42.02	51.90	72.04	**58.26**	**71.36**
Gemini 2.0 Flash	**74.29**	44.36	51.38	**83.99**	47.75	<u>72.84</u>	74.22	38.74	58.26	<u>75.90</u>	57.71	66.84
InternVL2.5-78B	<u>72.88</u>	**59.60**	**57.67**	80.95	<u>63.62</u>	**75.62**	<u>75.99</u>	<u>58.25</u>	**63.65**	**76.86**	56.20	<u>70.07</u>
LLaVA-Video-72B	53.11	54.94	53.14	65.34	60.45	67.03	67.94	52.39	53.49	70.66	56.20	70.01
InternVL2.5-26B	71.05	47.74	50.53	80.95	58.99	72.17	**76.13**	47.20	<u>60.12</u>	73.14	50.00	65.61
InternVL2.5-8B	69.77	<u>58.62</u>	49.96	80.95	**64.55**	71.02	73.94	55.80	54.66	72.18	51.79	68.19
VideoLLama3-8B	68.08	45.62	49.68	81.35	54.89	68.36	72.99	50.75	51.62	69.28	53.44	67.90
LLaVA-ov-7B	54.66	51.69	47.49	62.96	53.04	61.77	64.53	49.66	49.48	68.60	52.75	64.73
LLaVA-Video-7B	50.14	55.23	44.52	61.64	50.53	60.50	63.57	52.52	49.97	66.39	49.59	63.03
Qwen2-VL-7B	65.96	49.15	48.41	82.14	43.39	67.03	71.21	45.57	50.52	67.49	49.72	65.02
VILA-U-7B	58.19	39.83	38.87	63.10	41.93	54.51	59.07	37.93	37.94	59.50	33.88	55.34
VILA1.5-7B	54.80	55.93	39.29	66.93	62.30	63.52	55.25	**58.53**	44.64	63.64	52.34	61.91

Table 3: Performance of MLLMs on COVER, based on our quadrant formulation (A&C, C&C, C&P, A&P), measured by original, counterfactual, and sub-question accuracy.

Model	Without CoT	With CoT	Guide-CoT	
	cf_{acc}	cf_{acc}	cf_{acc}	$cf_{withans}$
GPT-4o-mini	51.47	<u>58.62</u>	<u>57.93</u>	<u>68.07</u>
InternVL2.5-78B	**59.46**	**60.42**	**58.33**	**68.29**
LlaVA-Video-72B	56.04	56.24	53.51	63.12
InternVL2.5-8B	<u>57.75</u>	57.06	52.41	57.75
VideoLlama3-8B	51.25	52.82	53.06	52.79
LLaVA-Video-7B	51.93	51.42	51.39	54.12
Qwen2-VL-7B	46.90	50.36	45.71	50.88

Table 4: Comparison between CoT and Guide-CoT performance across MLLMs on the COVER benchmark.

their smaller parameter size, revealing distinct capabilities across model types. Commercial models, such as the Gemini series, maintain strong performance in concrete cognition and abstract perception tasks but generally fall behind open-source models in counterfactual reasoning. Most models struggle with counterfactual reasoning, with only InternVL2.5-7BB and VILA1.5-7B showing some task-specific advantages, highlighting the need for targeted optimization in conditional hypothesis modeling.

4.3 Sub-question Guideline

We propose Guide-CoT to study the influence of different reasoning paths on model performance through human-annotated sub-problems. We design comparative experiments between CoT and Guide-CoT to analyze how automatically generated sub-questions from CoT versus manually annotated sub-questions affect model reasoning capabilities.

Comparing the Without CoT and CoT approaches based on Table 4, we find that the cf_{acc}

of most models under CoT significantly exceeds the Without CoT baseline, such as Qwen2-VL-7B (+3.46%) and GPT-4o-mini (+7.15%), which indicates that CoT enhances reasoning processes, particularly in more complex tasks.

However, examining Guide-CoT results reveals that manually designed sub-questions may not always lead to substantial improvement over automatically generated ones, as seen with GPT-4o-mini's cf_{acc} of 57.93% under Guide-CoT, slightly lower than the 58.62% under CoT. This does not imply the ineffectiveness of manual sub-questions but suggests that model behaviors may not always align with human-designed reasoning paths, potentially due to task complexity or the nature of the sub-questions themselves. We hypothesize that manually provided sub-questions could introduce extraneous patterns or "pseudo-features" that are not directly relevant to the reasoning task, leading to a subtle reduction in performance.

The $cf_{withans}$ column in Guide-CoT indicates sub-questions that include standard answers. For

Frames	InternVL2.5-1B			InternVL2.5-2B			InternVL2.5-4B			InternVL2.5-8B		
	ori_{acc}	cf_{acc}	sub_{acc}	ori_{acc}	cf_{acc}	sub_{acc}	ori_{acc}	cf_{acc}	sub_{acc}	ori_{acc}	cf_{acc}	sub_{acc}
2	66.16	35.61	55.27	65.31	44.20	54.99	72.56	48.31	60.88	71.26	58.50	60.07
4	68.32	34.72	55.52	68.83	42.11	58.84	74.41	46.49	61.79	73.35	58.47	60.96
8	68.94	35.10	55.11	68.22	41.43	55.75	75.03	45.60	61.79	74.14	57.06	61.60
16	69.76	35.89	55.19	70.07	40.68	55.49	75.61	45.23	61.63	74.31	57.75	61.65
32	69.04	36.50	55.04	70.13	39.69	55.48	75.54	45.09	60.96	74.10	57.03	61.42
64	68.18	37.39	54.80	68.90	40.06	55.44	74.41	46.56	60.70	74.20	58.09	61.30

Table 5: Performance of MLLMs on COVER using different frame sampling strategies. The frame selection follows standard practices in video QA benchmarks, where the number of input frames is set to $\min(\text{video length}, \text{predefined sampling count})$.

InternVL2.5-78B, $cf_{withans}$ reaches 68.29%, reflecting an 8.63% improvement over the no-CoT baseline, in contrast to CoT's modest gain of only 0.96% (from 59.46% to 60.42%). This suggests that providing complete answers substantially enhances reasoning accuracy, particularly in complex or multi-step tasks. Standard-answer sub-questions enable the model to better integrate information and verify intermediate reasoning steps, resulting in improved consistency and overall performance. Detailed case studies are presented in Appendix Figure 9 to further illustrate these findings and analyze the interplay between reasoning paths and task complexity.

The results from our experiments strongly support the notion that reasoning plays a pivotal role in model robustness and generalization. Our study extends these insights by demonstrating that multimodal models, especially in the context of video tasks, rely heavily on robust reasoning capabilities for effective generalization. The significant performance improvements observed with counterfactual reasoning and sub-question decomposition highlight that models' ability to handle complex, conditional, and dynamic contexts is crucial for their robustness, a finding not fully explored in prior research.

5 Analysis

In this chapter, we begin by analyzing the impact of video frame sampling rates on MLLMs' video understanding and reasoning abilities. We then proceed with an in-depth examination of MLLMs' robustness and logical reasoning performance.

5.1 Ablation Study of Video Frames

As shown in Table 5, as the parameter size of LLMs increases, there is a rising trend in ori_{acc}, cf_{acc}, and sub_{acc}. For instance, with 16 frames,

the InternVL2.5-1B model achieves ori_{acc}, cf_{acc}, and sub_{acc} of 69.76%, 35.89%, and 55.19% respectively. The InternVL2.5-2B scores 70.07%, 40.62%, and 55.49%, while the InternVL2.5-4B reaches 75.61%, 45.23%, and 61.68%, indicating that larger LLMs have enhanced capabilities in handling complex problems. Under the same vision tower settings, ori_{acc} shows a clear upward trend as the number of frames increases. For example, the InternVL2.5-8B's ori_{acc} rises from 71.26% at 2 frames to 74.20% at 64 frames. However, cf_{acc} tends to decrease with more frames. The InternVL2.5-2B's cf_{acc} drops from 44.20% at 2 frames to 40.06% at 64 frames. Models with more parameters generally perform better in ori_{acc}, cf_{acc}, and sub_{acc}, highlighting the significant role of LLMs in multimodal reasoning. Additionally, increasing visual information (by raising the frame count) can enhance ori_{acc}, but excessive visual information, especially in complex or counterfactual reasoning scenarios, may impair the model's reasoning ability, leading to a decline in cf_{acc}.

5.2 Robustness and Logical Reasoning in MLLMs

The ability of MLLMs to answer original questions serves as a key indicator of their overall understanding capabilities, while performance on sub-questions reveals single-step reasoning proficiency. A notable observation is the strong Pearson correlation between ori_{acc} and sub_{acc} reaches 0.836, indicating a strong connection between model understanding and reasoning capabilities. Furthermore, as shown in Figure 5, the correlation between sub_{acc} and cf_{acc} is 0.608. These moderately strong correlations indicate that a model's ability to comprehend original questions plays a fundamental role in enabling effective step-by-step reasoning. Similarly, the correlation between ori_{acc} and sub_{acc} suggests that models with a higher understanding
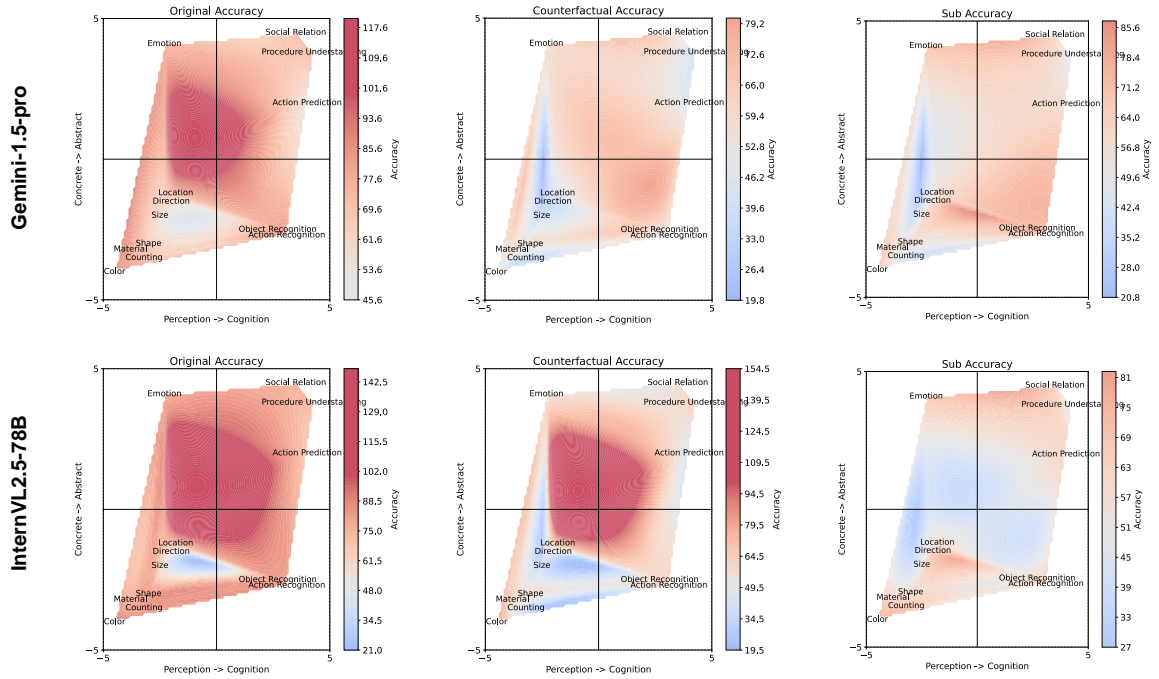
Figure 4: Heatmaps of task performance for Gemini-1.5-pro and InternVL2.5-78B, using hollow circles to depict task distributions across the four quadrants. The top three panels show results for Gemini-1.5-pro, and the bottom three for InternVL2.5-78B. **Left**: Accuracy on original questions. **Middle**: Performance on counterfactual questions. **Right**: Accuracy on sub-questions. A gradient color bar—from azure (low accuracy) to crimson (high accuracy)—is placed along the right margin of each heatmap to indicate performance levels.

Model	$P(cf_{right}\|sub_{right})$	$P(cf_{wrong}\|sub_{right})$	$P(cf_{right}\|sub_{wrong})$	$P(cf_{wrong}\|sub_{wrong})$
gemini-1.5-pro	56.54	**43.45**	44.99	**55.01**
GPT-4o-mini	59.49	40.51	47.65	52.35
InternVL2.5-78B	62.90	37.10	**56.67**	43.34
LlaVA-Video-72B	**63.28**	36.72	51.60	48.40

Table 6: Conditional probabilities of counterfactual accuracy given sub-question outcomes. $P(cf_{right} \mid sub_{right})$ and $P(cf_{wrong} \mid sub_{right})$ denote the likelihood of answering the counterfactual question correctly or incorrectly when the sub-question is correct; similarly, $P(cf_{right} \mid sub_{wrong})$ and $P(cf_{wrong} \mid sub_{wrong})$ apply when the sub-question is incorrect.

capability tend to perform better when solving decomposed sub-questions, reinforcing the notion that comprehension and reasoning are interdependent. However, the moderate correlation between sub_{acc} and cf_{acc} suggests that counterfactual reasoning involves additional complexities, making it a more challenging task than single-step reasoning.

As illustrated in Table 6, We observed that across multiple models, the probability $P(\text{cf\_right}|\text{sub\_right})$ was significantly higher than $P(\text{cf\_right}|\text{sub\_wrong})$, clearly indicating that the correctness of sub-questions is a strong predictor of overall counterfactual performance.

Analysis of the heat maps in Figure 4 reveals different performance patterns in the quadrants, high-lighting the interaction between comprehension, step-by-step reasoning, and counterfactual inference. In abstract reasoning tasks such as social inference and procedural understanding, the drop from sub_{acc} to ori_{acc} is minimal, and the transition to cf_{acc} remains stable. This suggests that models can effectively leverage sub-question reasoning and maintain performance even under counterfactual assumptions. In contrast, the **concrete perception quadrant**—involving tasks like object recognition and motion understanding—shows a sharper decline from sub_{acc} to ori_{acc}, and further to cf_{acc}. This indicates that perception-heavy tasks pose greater challenges, as models struggle to decompose complex sensory input into reasoning steps
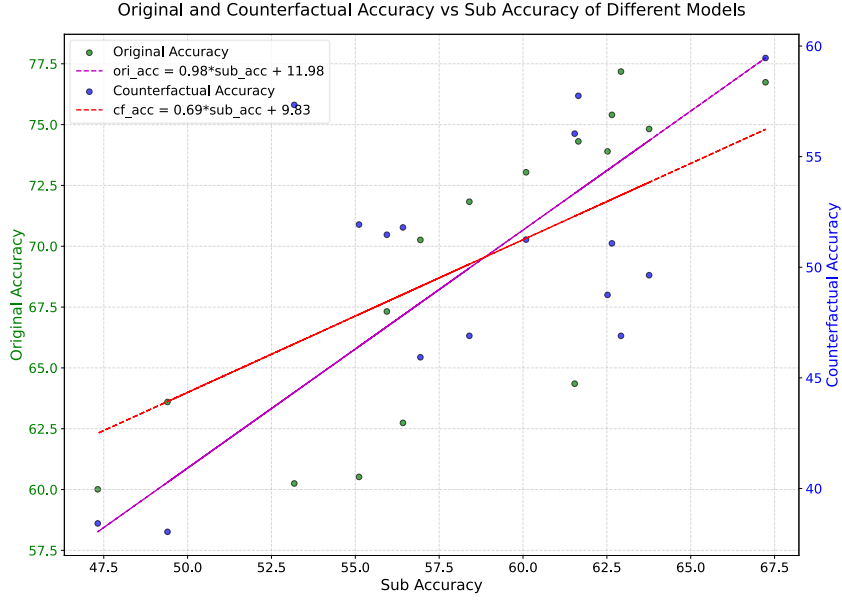
Figure 5: Scatter plot showing correlations among ori_{acc}, sub_{acc}, and cf_{acc} across models. The red line represents the linear function fitted between ori_{acc} and sub_{acc}, while the purple line represents the linear function fitted between cf_{acc} and sub_{acc}.

required for counterfactual understanding.

Overall, our findings indicate that counterfactual reasoning is inherently more challenging than single-step reasoning, especially in perception-intensive tasks where models must infer causality beyond pattern recognition. In contrast, the relatively stable gap between sub_{acc} and cf_{acc} in abstract-cognitive tasks suggests that models can better leverage conceptual knowledge. Enhancing counterfactual reasoning in perception-heavy scenarios remains a key challenge, likely requiring improved causal inference and reasoning mechanisms.

5.3 The Effects of Model Scale

We conduct systematic analyses to characterize performance gaps across original, counterfactual, and sub-question accuracies. Our goal is to mitigate these gaps by examining factors such as model scale, training alignment, and reasoning strategies. As shown in Table 7, with similar visual backbones, increasing language model size significantly reduces the performance gap—particularly between sub-question and counterfactual accuracy. Specifically, the absolute difference between ori_{acc} (70.07%) and cf_{acc} (40.68%) is 29.39% for the 2B model, increases slightly to 30.38% for the 4B

model, and then drops substantially to 16.56% for the 8B model. Similarly, the gap between cf_{acc} and sub_{acc} grows from 14.81% (2B) to 16.40% (4B), before narrowing sharply to 3.90% (8B).

Model	ori_{acc}	cf_{acc}	sub_{acc}
InternVL2.5-2B	70.07	40.68	55.49
InternVL2.5-4B	75.61	45.23	61.63
InternVL2.5-8B	74.31	57.75	61.65

Table 7: Variations in three accuracy metrics across different model sizes.

6 Conclusion

We introduce COVER, a comprehensive benchmark for counterfactual video reasoning that evaluates MLLMs across four dimensions: abstract-concrete and perception-cognition. By decomposing complex queries into structured sub-questions, COVER enables fine-grained analysis and reveals a strong correlation between sub-question accuracy and overall reasoning performance. Our results highlight the need for improved reasoning abilities in dynamic video tasks, and position COVER as a new standard for evaluating multimodal logical reasoning.

Acknowledgments

Limitations

COVER offers a novel benchmark for counterfactual video reasoning, but some limitations exist. First, while it focuses on video reasoning, its applicability to other multimodal tasks, such as image or text reasoning, remains unexplored. Second, COVER relies on sub-question decomposition, and automated methods may not always match human-designed questions, especially in complex scenarios. Finally, while we demonstrate COVER's effectiveness on various models, further validation across different model architectures and real-world tasks is needed to assess its generalizability.

Ethical Considerations

COVER is designed with ethical considerations in mind, aiming to enhance counterfactual reasoning in video understanding while ensuring fairness, transparency, and responsible AI development. We acknowledge the ongoing challenges in bias mitigation, fairness, and environmental sustainability and encourage the broader research community to collaborate in addressing these concerns. By establishing COVER as an open and structured evaluation benchmark, we aim to promote robust and ethical AI advancements in multimodal reasoning.

We ensured that the human annotators were compensated with fair remuneration, which exceeded the local minimum wage standards, reflecting the value of their work. Furthermore, we took steps to ensure that the annotation process did not pose any risks to their physical or mental well-being. The tasks were designed to be manageable, and we provided adequate support to ensure a safe and respectful working environment.

In this study, AI was used solely for data augmentation and grammar/typo correction, with no involvement in generative or creative tasks. We carefully considered potential risks to ensure AI usage did not compromise the originality or transparency of the research.

References

AI Anthropic. 2024. Claude 3.5 sonnet model card addendum. *Claude-3.5 Model Card*.

Guangsheng Bao, Hongbo Zhang, Cunxiang Wang, Linyi Yang, and Yue Zhang. 2025. How likely do LLMs with CoT mimic human reasoning? In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 7831–7850, Abu Dhabi, UAE. Association for Computational Linguistics.

Wenhao Chai, Enxin Song, Yilun Du, Chenlin Meng, Vashisht Madhavan, Omer Bar-Tal, Jenq-Neng Hwang, Saining Xie, and Christopher D. Manning. 2024. Auroracap: Efficient, performant video detailed captioning and a new benchmark. *CoRR*, abs/2410.03051.

Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, Lixin Gu, Xuehui Wang, Qingyun Li, Yimin Ren, Zixuan Chen, Jiapeng Luo, Jiahao Wang, Tan Jiang, Bo Wang, Conghui He, Botian Shi, Xingcheng Zhang, Han Lv, Yi Wang, Wenqi Shao, Pei Chu, Zhongying Tu, Tong He, Zhiyong Wu, Huipeng Deng, Jiaye Ge, Kai Chen, Min Dou, Lewei Lu, Xizhou Zhu, Tong Lu, Dahua Lin, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *CoRR*, abs/2412.05271.

Chaoyou Fu, Peixian Chen, Yunhang Shen, Yulei Qin, Mengdan Zhang, Xu Lin, Zhenyu Qiu, Wei Lin, Jinrui Yang, Xiawu Zheng, Ke Li, Xing Sun, and Rongrong Ji. 2023. MME: A comprehensive evaluation benchmark for multimodal large language models. *CoRR*, abs/2306.13394.

Chaoyou Fu, Yuhan Dai, Yondong Luo, Lei Li, Shuhuai Ren, Renrui Zhang, Zihan Wang, Chenyu Zhou, Yunhang Shen, Mengdan Zhang, Peixian Chen, Yanwei Li, Shaohui Lin, Sirui Zhao, Ke Li, Tong Xu, Xiawu Zheng, Enhong Chen, Rongrong Ji, and Xing Sun. 2024. Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. *CoRR*, abs/2405.21075.

Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. 2017. TALL: temporal activity localization via language query. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 5277–5285. IEEE Computer Society.

Drew A. Hudson and Christopher D. Manning. 2019. GQA: A new dataset for real-world visual reasoning and compositional question answering. In *IEEE*

Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019, pages 6700–6709. Computer Vision Foundation / IEEE.

Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Madry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll L. Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, and Dane Sherburn. 2024. Gpt-4o system card. *CoRR*, abs/2410.21276.

Yunseok Jang, Yale Song, Youngjae Yu, Youngjin Kim, and Gunhee Kim. 2017. TGIF-QA: toward spatio-temporal reasoning in visual question answering. In *2017 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2017, Honolulu, HI, USA, July 21-26, 2017*, pages 1359–1367. IEEE Computer Society.

Katja Katja Wiemer-Hastings and Xu Xu. 2005. Content differences for abstract and concrete concepts. *Cognitive science*, 29(5):719–736.

Jacob Krantz, Erik Wijmans, Arjun Majumdar, Dhruv Batra, and Stefan Lee. 2020. Beyond the nav-graph: Vision-and-language navigation in continuous environments. In *Computer Vision - ECCV 2020 - 16th European Conference, Glasgow, UK, August 23-28, 2020, Proceedings, Part XXVIII*, volume 12373 of *Lecture Notes in Computer Science*, pages 104–120. Springer.

Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li. 2024a. Llava-onevision: Easy visual task transfer. *CoRR*, abs/2408.03326.

Bohao Li, Rui Wang, Guangzhi Wang, Yuying Ge, Yixiao Ge, and Ying Shan. 2023. Seed-bench: Benchmarking multimodal llms with generative comprehension. *CoRR*, abs/2307.16125.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Lou, Limin Wang, and Yu Qiao. 2024b. Mvbench: A comprehensive multi-modal video understanding benchmark. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 22195–22206. IEEE.

Shicheng Li, Lei Li, Yi Liu, Shuhuai Ren, Yuanxin Liu, Rundong Gao, Xu Sun, and Lu Hou. 2024c. VITATECS: A diagnostic dataset for temporal concept understanding of video-language models. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXX*, volume 15128 of *Lecture Notes in Computer Science*, pages 331–348. Springer.

Yian Li, Wentao Tian, Yang Jiao, Jingjing Chen, and Yu-Gang Jiang. 2024d. Eyes can deceive: Benchmarking counterfactual reasoning abilities of multi-modal large language models. *CoRR*, abs/2404.12966.

Yian Li, Wentao Tian, Yang Jiao, Jingjing Chen, Na Zhao, and Yu-Gang Jiang. 2024e. Look before you decide: Prompting active deduction of mllms for assumptive reasoning. *Preprint*, arXiv:2404.12966.

Ji Lin, Hongxu Yin, Wei Ping, Pavlo Molchanov, Mohammad Shoeybi, and Song Han. 2024. VILA: on pre-training for visual language models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 26679–26689. IEEE.

Yuan Liu, Haodong Duan, Yuanhan Zhang, Bo Li, Songyang Zhang, Wangbo Zhao, Yike Yuan, Jiaqi Wang, Conghui He, Ziwei Liu, Kai Chen, and Dahua Lin. 2024. Mmbench: Is your multi-modal model an all-around player? In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part VI*, volume 15064 of *Lecture Notes in Computer Science*, pages 216–233. Springer.

Maitreya Patel, Tejas Gokhale, Chitta Baral, and Yezhou Yang. 2022. CRIPP-VQA: counterfactual reasoning about implicit physical properties via video question answering. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 9856–9870. Association for Computational Linguistics.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, Yusuf Aytar, Simon

Osindero, Dima Damen, Andrew Zisserman, and João Carreira. 2023. Perception test: A diagnostic benchmark for multimodal video models. *Preprint*, arXiv:2305.13786.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy P. Lillicrap, Jean-Baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew M. Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, Yuanzhong Xu, James Molloy, Jilin Chen, Michael Isard, Paul Barham, Tom Hennigan, Ross McIlroy, Melvin Johnson, Johan Schalkwyk, Eli Collins, Eliza Rutherford, Erica Moreira, Kareem Ayoub, Megha Goel, Clemens Meyer, Gregory Thornton, Zhen Yang, Henryk Michalewski, Zaheer Abbas, Nathan Schucher, Ankesh Anand, Richard Ives, James Keeling, Karel Lenc, Salem Haykal, Siamak Shakeri, Pranav Shyam, Aakanksha Chowdhery, Roman Ring, Stephen Spencer, Eren Sezener, and et al. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *CoRR*, abs/2403.05530.

Amir Shahroudy, Jun Liu, Tian-Tsong Ng, and Gang Wang. 2016. NTU RGB+D: A large scale dataset for 3d human activity analysis. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pages 1010–1019. IEEE Computer Society.

Gunnar A. Sigurdsson, Gül Varol, Xiaolong Wang, Ali Farhadi, Ivan Laptev, and Abhinav Gupta. 2016. Hollywood in homes: Crowdsourcing data collection for activity understanding. In *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, volume 9905 of *Lecture Notes in Computer Science*, pages 510–526. Springer.

Ganchao Tan, Daqing Liu, Meng Wang, and Zheng-Jun Zha. 2020. Learning to discretely compose reasoning module networks for video captioning. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 745–752. ijcai.org.

Qwen Team. 2024. Qvq: To see the world with wisdom.

Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. 2024. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. 2019. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pages 4580–4590. IEEE.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Jian Wu, Linyi Yang, Zhen Wang, Manabu Okumura, and Yue Zhang. 2024a. Cofca: A step-wise counterfactual multi-hop qa benchmark. *Preprint*, arXiv:2402.11924.

Te-Lin Wu, Zi-Yi Dou, Qingyuan Hu, Yu Hou, Nischal Chandra, Marjorie Freedman, Ralph Weischedel, and Nanyun Peng. 2023. ACQUIRED: A dataset for answering counterfactual questions in real-life videos. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 11753–11770, Singapore. Association for Computational Linguistics.

Yecheng Wu, Zhuoyang Zhang, Junyu Chen, Haotian Tang, Dacheng Li, Yunhao Fang, Ligeng Zhu, Enze Xie, Hongxu Yin, Li Yi, Song Han, and Yao Lu. 2024b. VILA-U: a unified foundation model integrating visual understanding and generation. *CoRR*, abs/2409.04429.

Zhaofeng Wu, Linlu Qiu, Alexis Ross, Ekin Akyürek, Boyuan Chen, Bailin Wang, Najoung Kim, Jacob Andreas, and Yoon Kim. 2024c. 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, Mexico City, Mexico. Association for Computational Linguistics.

Binzhu Xie, Sicheng Zhang, Zitang Zhou, Bo Li, Yuanhan Zhang, Jack Hessel, Jingkang Yang, and Ziwei Liu. 2024. [inline-graphic not available: see fulltext] funqa: Towards surprising video comprehension. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part I*, volume 15059 of *Lecture Notes in Computer Science*, pages 39–57. Springer.

Linyi Yang, Shuibai Zhang, Libo Qin, Yafu Li, Yidong Wang, Hanmeng Liu, Jindong Wang, Xing Xie, and Yue Zhang. 2023. GLUE-X: Evaluating natural language understanding models from an out-of-distribution generalization perspective. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12731–12750, Toronto, Canada. Association for Computational Linguistics.

Kexin Yi, Chuang Gan, Yunzhu Li, Pushmeet Kohli, Jiajun Wu, Antonio Torralba, and Joshua B. Tenenbaum. 2020. Clevrer: Collision events for video representation and reasoning. *Preprint*, arXiv:1910.01442.

Weihao Yu, Zhengyuan Yang, Linjie Li, Jianfeng Wang, Kevin Lin, Zicheng Liu, Xinchao Wang, and Lijuan

Wang. 2024. Mm-vet: Evaluating large multimodal models for integrated capabilities. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, Peng Jin, Wenqi Zhang, Fan Wang, Lidong Bing, and Deli Zhao. 2025. Videollama 3: Frontier multimodal foundation models for image and video understanding. *CoRR*, abs/2501.13106.

Hongjie Zhang, Yi Liu, Lu Dong, Yifei Huang, Zhen-Hua Ling, Yali Wang, Limin Wang, and Yu Qiao. 2023. Movqa: A benchmark of versatile question-answering for long-form movie understanding. *Preprint*, arXiv:2312.04817.

Yuanhan Zhang, Jinming Wu, Wei Li, Bo Li, Zejun Ma, Ziwei Liu, and Chunyuan Li. 2024a. Video instruction tuning with synthetic data. *CoRR*, abs/2410.02713.

Zhuosheng Zhang, Aston Zhang, Mu Li, Hai Zhao, George Karypis, and Alex Smola. 2024b. Multimodal chain-of-thought reasoning in language models. *Preprint*, arXiv:2302.00923.

Ge Zheng, Bin Yang, Jiajin Tang, Hong-Yu Zhou, and Sibei Yang. 2023. Ddcot: Duty-distinct chain-of-thought prompting for multimodal reasoning in language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*.

A Appendix

A.1 Data Construction Details

In this section, we present additional details on COVER construction, including information about the task splitting scores, annotation agreements, data augmentation prompts and process flow.

We invited three expert annotators to independently score each benchmark task based on our two-dimensional quadrant framework (abstract vs. concrete and perception vs. cognition). Their scoring results in Table 8 demonstrates the strictness, consistency, and logical coherence of our task categorization, effectively preventing overlaps and ambiguity.

The annotators were recruited to evaluate COVER across multiple dimensions, with the resultant assessments systematically compiled in Table 9, ensuring comprehensive evaluation coverage. The methodological workflow for data augmentation is schematically outlined in Figure 6.

The schematic framework outlined in Figure 7 delineates the methodology employed for contextual data augmentation, leveraging the generative capabilities of GPT-4o(Hurst et al., 2024) to construct domain-specific instructional prompts.

A.2 Additional Results

In this section, we present additional experiments on COVER. The comprehensive evaluation framework delineated in Table 14 presents granular performance metrics across 13 meticulously defined tasks.

GPT-4o exhibited notable vulnerability in the Procedure Understanding task. While it attained a respectable raw accuracy of 78.17%, its counterfactual accuracy plummeted to 28.97%, representing a precipitous decline of 49.2%. This substantial drop suggests that the performance of GPT-4o in understanding procedures may be overly reliant on surface-level features. Counterfactual perturbations, such as changes in conditions, can severely disrupt its reasoning capabilities, thereby highlighting a robustness limitation of the model when handling complex tasks.

Figure 5 (a) depicts the relationship between ori_{acc} and sub_{acc} across different models, with a purple regression line characterizing the functional correlation between mean ori_{acc} and mean sub_{acc}. Figure 5 (b) demonstrates the association between cf_{acc} and sub_{acc} across different models, with a red regression line characterizing the functional correlation between mean cf_{acc} and mean sub_{acc}. The bivariate correlation analysis delineated in Figure 5 demonstrates statistically significant covariation patterns (r = 0.836) between semantic comprehension and multi-step reasoning capabilities in MLLMs.

We conducted an additional ablation study to examine whether the observed trend where excessive visual information impairs reasoning accuracy holds consistently across both short and long videos. Our results are summarized in Table 10, 11. We observed a clear pattern across both short and long videos: model accuracy typically peaks within a moderate frame range (8–32 frames) and subsequently declines at the maximum setting (64 frames). This decline is particularly pronounced in tasks involving the original questions (ori) and sub questions (sub), suggesting that an excessive amount of visual input can indeed negatively impact model performance, regardless of video length.

Task	A_x	A_y	B_x	B_y	C_x	C_y	Avg$_x$	Avg$_y$
Counting	-3.2	-3.4	-3.1	-3.6	-3.3	-3.7	-3.2	-3.57
Color	-4.1	-4.4	-4.4	-4.2	-4.2	-4.3	-4.23	-4.3
Material	-3.8	-3.3	-3.9	-3.2	-4.0	-3.4	-3.9	-3.3
Size	-2.4	-2.5	-2.6	-2.3	-2.2	-2.4	-2.4	-2.4
Shape	-3.3	-3.2	-3.5	-3.2	-3.8	-4.0	-3.53	-3.47
Emotion	-2.4	4.0	-2.5	3.5	-2.4	3.1	-2.43	3.53
Location	-1.7	-1.4	-2.0	-1.6	-1.3	-1.7	-1.67	-1.57
Direction	-2.1	-1.7	-2.5	-1.5	-2.6	-1.8	-2.4	-1.67
Object Recognition	3.0	-3.0	2.4	-2.0	1.2	-2.3	2.2	-2.43
Action Recognition	2.5	-3.1	2.3	-3.0	2.1	-3.5	2.3	-3.2
Action Prediction	3.9	2.4	3.8	2.5	3.2	2.2	3.63	2.37
Procedure Understanding	3.0	3.5	3.6	3.2	2.2	3.3	2.93	3.33
Social Relation	3.4	4.3	3.0	4.4	3.1	4.1	3.17	4.27

Table 8: Annotator scoring table. Annotators A, B, and C provide ratings along two axes: the perceptual–cognitive dimension (x-axis, from -5 to 5, where higher values indicate more cognitive tasks) and the concrete–abstract dimension (y-axis, from -5 to 5, where higher values indicate more abstract tasks).
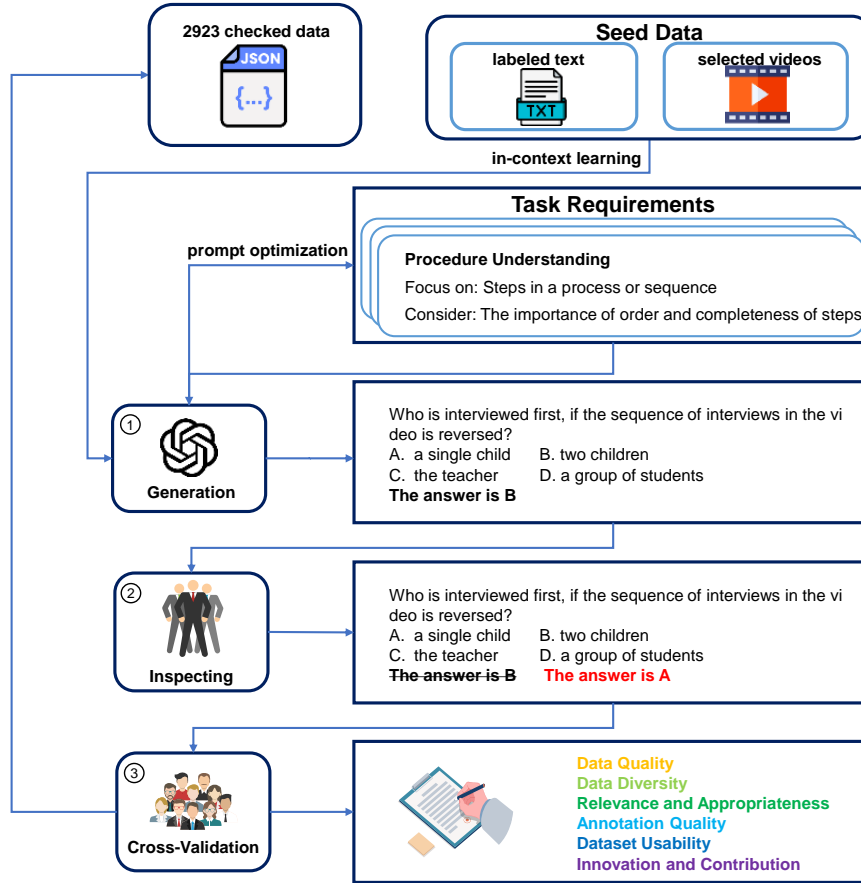
Figure 6: Flowchart depicting the data augmentation pipeline.

Additionally, we evaluated test-time reasoning strategies on manually curated seed data using long-chain reasoning models in Table 12. Notably, models such as InternVL2.5-78B-CoT show significant improvement in bridging the cf–sub–ori gap, further supporting that reasoning-guided prompting (e.g., CoT) helps align sub-level and cf-level accu-racy. These observations suggest a promising direction: larger and better-aligned models, when combined with explicit reasoning strategies, are more capable of maintaining coherence across perception, decomposition, and abstract reasoning tasks.

Figure 7: Methodological framework for data augmentation using GPT-4o.

Aspect	A	B	C	Average
Data Quality	4	4	5	4.3
Data Diversity	5	4	5	4.7
Relevance	4	5	4	4.3
Annotation Quality	4	5	5	4.7
Dataset Usability	4	4	4	4
Innovation	5	5	4	4.7

Table 9: Cross-annotator validation on COVER. The table summarizes quality scores assigned by three annotators. A, B, and C denote randomly assigned codes for the assessment data, and *Average* indicates the mean score across all entries.

Frame	InternVL2.5-4B			InternVL2.5-8B		
	ori_{acc}	cf_{acc}	sub_{acc}	ori_{acc}	cf_{acc}	sub_{acc}
2	69.09	45.94	60.33	68.72	56.53	61.14
4	70.81	46.18	**60.91**	68.97	56.90	60.68
8	**71.31**	43.97	60.62	69.83	56.28	61.14
16	70.81	44.83	59.86	**70.07**	56.40	61.20
32	70.69	43.84	59.57	69.21	56.90	**61.26**
64	69.95	**46.55**	59.63	69.09	**57.27**	60.62

Table 10: Performance of MLLMs with different numbers of sampled frames for short videos (1–64 frames).

A.3 Sample Reaults on Test Time Long Reasoning Models

As illustrated in Figure 10, the reasoning model QVQ-72B-Preview (Team, 2024), equipped with a built-in Chain-of-Thought (CoT) mechanism, exhibits human-aligned reasoning patterns. Its cognitive process integrates detailed scenario descriptions, systematic elimination of implausible options (e.g., excluding candidates A/B/C), and rigorous conclusion verification. In contrast, InternVL2.5-78B employs a CoT mechanism that presents an-

Frame	InternVL2.5-4B			InternVL2.5-8B		
	ori_{acc}	cf_{acc}	sub_{acc}	ori_{acc}	cf_{acc}	sub_{acc}
2	73.90	**49.22**	61.09	72.24	**59.26**	59.67
4	75.79	46.61	62.13	75.04	59.07	61.07
8	76.46	46.23	62.24	75.79	57.37	61.78
16	**77.45**	45.38	**62.31**	75.94	58.27	**61.82**
32	77.40	45.57	61.49	75.98	57.08	61.49
64	76.13	46.57	61.11	**76.17**	58.41	61.55

Table 11: Effect of different frame sampling strategies on MLLM performance for long videos (64–2000 frames).

Model	ori_acc	cf_acc	sub_acc
QVQ-72B-Preview	69.33	59.33	58.76
InternVL2.5-78B-CoT	70.00	71.33	70.80

Table 12: Variation in accuracy across different test-time reasoning strategies.

swers in a bullet-point format without explanatory justification, reflecting weaker anthropomorphic reasoning characteristics.

However, the cf_{acc} discrepancy in Table 13 (QVQ-72B-Preview: 59.33% < InternVL2.5-78B: 71.33%) suggests that contemporary reasoning models may rely more on memorization than on structured reasoning. InternVL2.5-78B's concise response paradigm appears to leverage rapid pattern recognition and information retrieval, leading to superior accuracy. While QVQ-72B-Preview's elaborate reasoning workflow better approximates human cognition, potential redundancies or logical inconsistencies may reduce answer precision.

Table 13 further indicates that InternVL2.5-78B achieves a substantial lead in the sub_{acc} metric (70.80%), significantly outperforming QVQ-72B-Preview (58.76%) and Claude-3.7-sonnect (46.72%). This performance hierarchy remains consistent across models when evaluated on the ori_{acc} metric: InternVL2.5-78B (70.00%) > QVQ-72B-Preview (69.33%) > Claude-3.7-sonnect (46.00%). Empirical evidence suggests a statistically significant positive correlation between reasoning capability (sub_{acc}) and comprehension ability (ori_{acc}). In addition, under the CoT paradigm, reasoning capability demonstrates stronger generalization, exhibiting a positive correlation with performance on human-annotated essential logical sub-problems, thereby reinforcing the intrinsic relationship between logical reasoning and generalizability.

Moreover, the reasoning processes of models such as QVQ frequently generate sub-problem content that aligns with human-annotated data, which to some extent suggests that the inferential patterns of test-time long-reasoning models demonstrate closer correspondence with human cognitive intuition. For instance, in the Figure 11 the analytical content regarding the opening and closing scenes of videos (highlighted in blue font) exhibits precise alignment with the manually curated sub-problems in the upper-right annotation (specifically addressing inquiries about video commencement and conclusion scenarios), thereby empirically validating

	ori_{acc}	cf_{acc}	sub_{acc}
QVQ-72B-Preview	69.33	59.33	58.76
Claude-3.7-sonnect	46.00	59.33	46.72
InternVL2.5-78B	70.00	71.33	70.80
VILA1.5-13B	65.33	44.67	53.65

Table 13: Performance of different chain-of-thought (CoT) reasoning architectures on a manually annotated dataset of 150 samples. QVQ and Claude-3.5-Sonnet represent dedicated reasoning models, while the others apply CoT-based augmentation.

this cognitive congruence.

A.4 Examples of Sub-question Guidelines

Figure 8 illustrates how sub-question errors propagate to counterfactual question failures. In Figure 9, we observe that subtle errors in the reasoning process lead to reasoning failures, highlighting the model's sensitivity to the integrity of its reasoning steps.

Model	Type	Task												
		Action Prediction	*Procedure Understanding*	*Social Relation*	*Action Recognition*	*Object Recognition*	Color	Counting	Direction	Location	Material	Shape	Size	Emotion
GPT-4o	ori_{acc}	65.20	78.17	69.00	74.87	74.87	92.23	75.25	50.88	70.59	79.12	72.00	52.88	65.01
	cf_{acc}	41.41	28.97	56.33	44.65	42.67	37.86	40.59	33.33	42.86	59.34	58.00	29.81	55.65
	sub_{acc}	51.85	22.82	52.43	69.54	67.09	51.94	47.52	56.90	48.96	58.08	55.56	36.08	63.97
GPT-4o-mini	ori_{acc}	50.22	72.22	63.32	78.61	74.08	84.47	70.30	52.63	57.98	71.43	62.00	56.73	65.56
	cf_{acc}	44.05	51.19	62.01	56.42	52.36	26.21	39.60	36.84	59.66	52.75	53.00	17.31	58.26
	sub_{acc}	53.16	19.44	58.85	68.03	64.82	38.35	38.12	53.88	47.72	53.89	53.44	28.85	65.96
Claude-3.5-Sonnet	ori_{acc}	43.61	63.10	63.32	66.31	73.82	79.61	68.32	48.25	52.10	63.74	66.34	45.19	66.94
	cf_{acc}	39.21	33.33	38.86	43.85	40.84	36.89	19.80	35.96	40.34	37.36	40.59	18.27	39.81
	sub_{acc}	46.19	15.87	46.68	62.54	60.93	39.81	24.26	48.28	48.55	46.11	37.70	34.13	56.81
Gemini-1.5-Pro	ori_{acc}	54.63	80.95	71.62	83.42	80.89	84.47	73.27	61.40	76.47	81.32	67.33	59.62	75.48
	cf_{acc}	46.70	29.76	58.52	45.45	58.38	46.60	37.62	35.96	42.86	54.95	55.45	36.54	57.99
	sub_{acc}	57.52	39.68	64.38	72.58	72.99	73.30	43.56	61.21	58.09	59.88	55.50	45.67	69.54
Gemini-1.5-Flash	ori_{acc}	53.74	85.32	70.74	82.62	81.41	82.52	70.30	57.02	70.59	79.12	69.31	68.27	72.04
	cf_{acc}	45.81	34.92	56.33	49.20	49.48	41.75	37.62	33.33	41.18	64.84	53.47	25.96	58.26
	sub_{acc}	61.87	32.94	63.94	73.28	69.60	46.60	43.07	62.93	54.77	62.87	55.50	37.98	71.36
Gemini-2.0-Flash	ori_{acc}	60.35	86.90	74.24	87.97	80.10	90.29	69.31	64.04	78.99	81.32	70.30	66.35	75.90
	cf_{acc}	42.29	36.51	51.97	44.12	51.31	20.39	39.60	31.58	37.82	57.14	56.44	31.73	57.71
	sub_{acc}	60.78	35.52	59.51	73.16	72.49	69.90	59.41	65.95	53.53	58.08	58.64	42.31	66.84
InternVL2.5-78B	ori_{acc}	67.84	75.00	75.55	79.68	82.20	94.17	82.18	52.63	76.47	76.92	83.17	69.23	76.86
	cf_{acc}	43.61	76.19	57.21	65.51	61.78	87.38	37.62	47.37	75.63	61.54	57.43	39.43	56.20
	sub_{acc}	62.09	44.64	67.70	76.90	62.28	79.13	69.80	66.38	58.09	62.28	59.69	50.48	70.07
LLaVA-Video-72B	ori_{acc}	43.17	50.79	65.50	60.70	69.90	85.44	69.31	51.75	73.11	74.73	61.39	61.54	70.66
	cf_{acc}	44.93	59.92	59.39	63.10	57.85	62.14	42.57	47.37	66.39	53.85	51.49	41.35	56.20
	sub_{acc}	59.26	32.94	69.47	67.56	66.46	63.59	52.97	61.21	45.23	55.69	53.40	43.27	70.01
InternVL2.5-26B	ori_{acc}	57.27	78.58	76.42	82.35	79.58	91.26	74.26	62.28	85.71	74.73	78.22	66.35	73.14
	cf_{acc}	47.14	45.24	51.09	60.43	57.59	59.23	25.74	45.61	60.50	57.14	25.00	25.00	50.00
	sub_{acc}	59.91	61.08	62.39	71.18	73.24	65.05	56.44	68.97	58.09	61/08	50.96	50.96	65.61
InternVL2.5-8B	ori_{acc}	55.51	75.00	78.17	81.28	80.63	90.29	70.30	63.16	78.99	74.63	74.26	66.35	72.18
	cf_{acc}	48.02	76.19	49.78	71.39	57.85	84.47	36.63	53.51	70.59	59.34	55.45	28.85	51.79
	sub_{acc}	55.99	29.37	66.81	69.89	72.24	52.43	52.97	60.34	53.53	56.89	54.97	51.44	68.19
VideoLLama3-8B	ori_{acc}	52.42	81.75	68.56	80.48	82.20	94.17	70.30	63.16	81.51	70.33	68.32	62.50	69.28
	cf_{acc}	35.68	53.97	46.29	55.08	54.71	66.02	42.57	42.11	64.71	58.24	48.51	32.69	53.44
	sub_{acc}	49.45	33.93	67.48	67.91	68.84	57.77	39.60	58.62	49.79	53.89	54.45	47.12	67.90
LLaVA-ov-7B	ori_{acc}	48.90	48.81	66.81	60.43	65.45	86.41	63.37	44.74	63.03	72.53	61.39	63.46	68.60
	cf_{acc}	43.61	64.29	45.85	55.35	50.79	59.22	42.57	45.61	52.94	60.44	57.43	30.77	52.75
	sub_{acc}	50.11	30.36	63.94	62.78	60.68	54.85	42.08	53.45	45.64	50.90	48.69	50.96	64.73
LLaVA-Video-7B	ori_{acc}	50.66	35.71	65.50	56.15	67.02	83.50	64.36	41.23	58.82	74.72	64.36	59.62	66.39
	cf_{acc}	44.93	73.02	45.85	59.09	42.15	73.79	42.57	48.25	60.50	58.24	48.51	35.58	49.59
	sub_{acc}	48.58	29.96	56.64	62.19	58.67	55.34	41.09	56.03	43.15	54.49	52.88	48.08	63.03
Qwen2-VL-7B	ori_{acc}	44.49	84.12	67.25	84.22	80.10	88.35	70.29	57.89	73.94	74.73	65.34	69.23	67.49
	cf_{acc}	42.29	58.73	45.41	44.92	41.88	56.31	21.78	42.11	58.82	60.44	45.54	33.65	49.72
	sub_{acc}	53.37	30.16	63.72	67.33	66.71	43.20	42.08	59.48	51.87	52.69	51.83	51.44	65.02
VILA-U-7B	ori_{acc}	45.37	73.02	54.59	66.31	59.95	81.55	61.39	54.39	71.43	47.25	47.52	47.11	59.50
	cf_{acc}	22.47	53.17	42.36	44.39	39.53	45.63	45.54	23.68	43.70	35.16	35.64	36.54	33.88
	sub_{acc}	38.78	34.33	44.03	52.16	57.04	41.26	22.28	39.22	40.66	35.93	42.93	42.31	55.34
VILA1.5-7B	ori_{acc}	52.86	50.40	61.57	67.65	66.23	61.17	56.44	40.35	40.34	71.43	60.40	62.50	63.64
	cf_{acc}	28.64	85.71	50.22	68.45	56.29	86.41	57.43	49.12	76.47	51.65	41.58	44.23	52.34
	sub_{acc}	34.86	24.80	59.96	61.73	65.45	48.06	31.19	50.86	42.74	47.31	45.55	46.63	61.91

Table 14: Overall performance of MLLMs on 13 tasks in COVER, including original accuracy, counterfactual accuracy, and sub-question accuracy.

2955
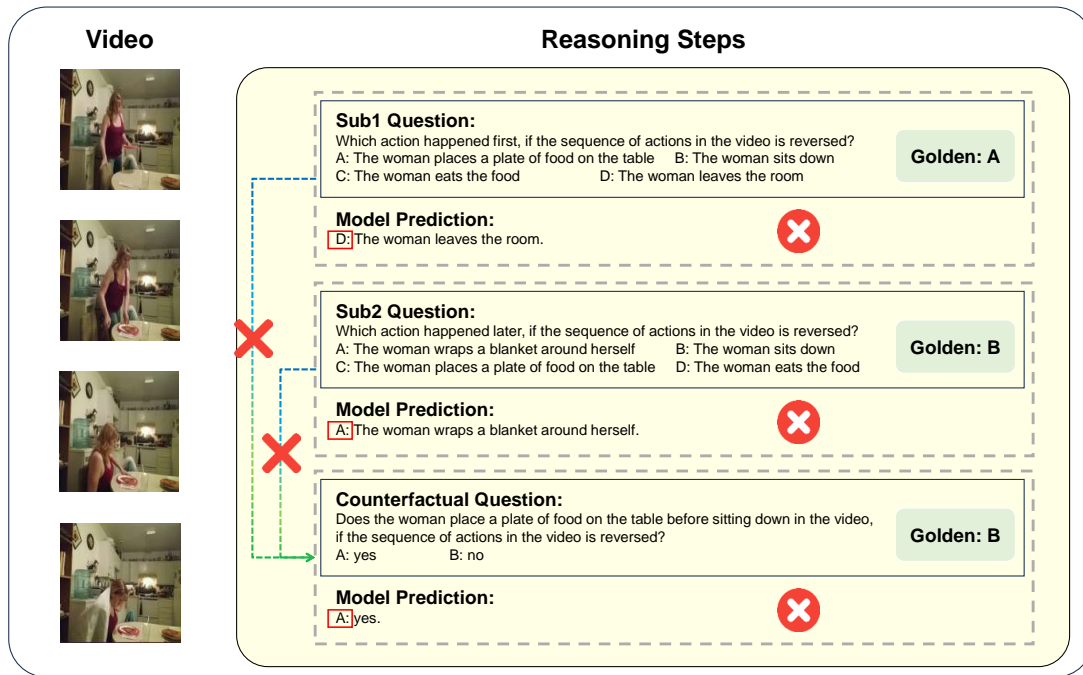
Figure 8: Example from **COVER**, showing a video accompanied by three related questions. The video is divided into four key action frames (left), with dashed lines indicating reasoning steps. Single-step prediction errors are marked with red crosses on the right, while sub-questions that do not support counterfactual reasoning are marked with red crosses on the left.
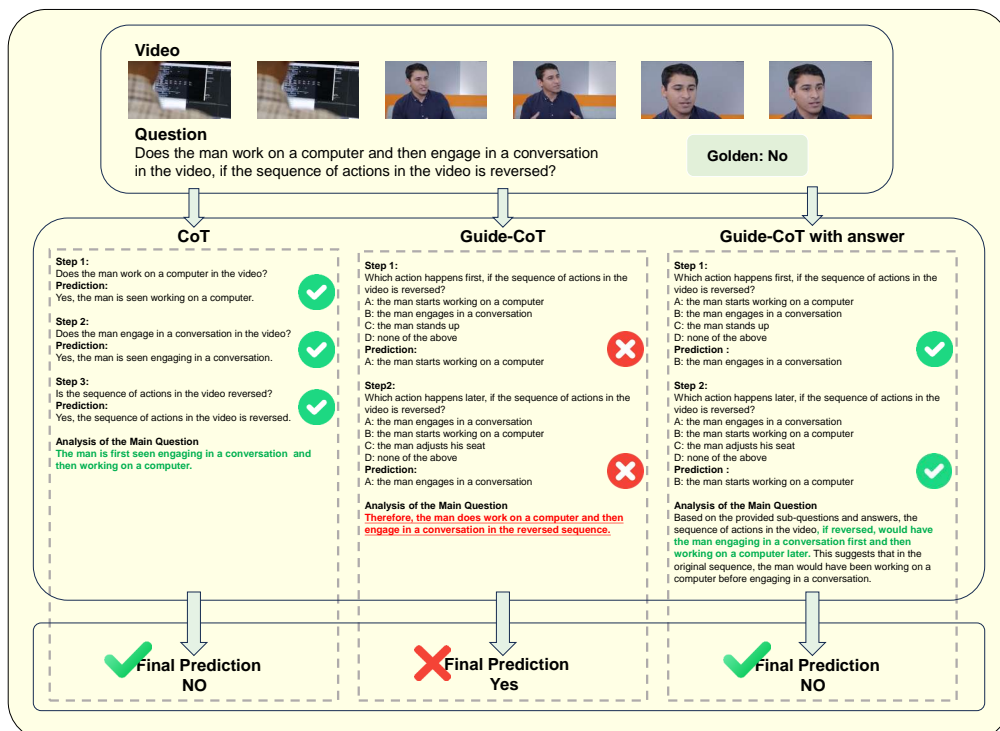
Figure 9: An example from **COVER**. The top section shows the video input and corresponding counterfactual questions. The middle section presents three reasoning processes—CoT, Guide-CoT, and Standard—where correct steps are marked with green checkmarks. In the analysis, correct reasoning paths are shown in green text, while incorrect ones are highlighted in red. The bottom section displays the final model predictions, with green checkmarks indicating correct answers and red crosses denoting errors.

2956

Video

Question
Is the woman leaving or returning home in the video, if the video is played in reverse?
A: leaving   B: sitting in the car   C: walking in the park   D: returning

Golden
D: returning

QVQ

let's start by describing what I see.
First, there's a woman in a white tank top and blue jeans. She's walking out of a house through a blue door. The house looks like a single-story building with white siding. As she exits, she walks down a small set of stairs and heads towards a driveway where several vehicles are parked. In the driveway, there's a black car parked closest to the house, followed by a dark blue pickup truck, and then another black vehicle that looks like a larger truck or possibly a tractor trailer. The woman walks past these vehicles, seemingly in a hurry, as she moves quickly across the driveway and onto a grassy area beyond. The setting appears to be rural or semi-rural, given the open spaces, the style of the house, and the types of vehicles present. There are some trees and other buildings visible in the background, further suggesting a countryside environment.
Now, the question mentions that if the video is played in reverse, the woman would be leaving or returning home. The options provided are: A: leaving B: sitting in the car C: walking in the park D: returning. Given that in the original video, the woman is seen exiting the house and walking away from it, playing the video in reverse would show her walking towards the house and entering through the blue door. So, in reverse, she would be returning home. **Therefore, the correct answer should be D: returning.**

But let's double-check to make sure I'm not missing anything.
The question says, "if the video is played in reverse," and the woman is seen walking out of the house in the original video. Reversing the video would show her walking back into the house, which implies she's returning home. Option A: leaving \u2013 that would correspond to the original direction, not the reverse. Option B: sitting in the car \u2013 there's no indication of her sitting in the car in the reverse video. Option C: walking in the park \u2013 while the reverse video might show her walking, it's not specifically in a park; it's more likely the same path back to the house. Option D: returning \u2013 this seems the most accurate based on the reverse scenario. **So, my final answer is D: returning.**

InternVL2.5-CoT

Step 1:
What is the woman doing at the beginning of the video?
Prediction :
The woman is putting on her boots ✅

Step 2:
What is the woman doing at the end of the video?
Prediction :
The woman is walking towards the truck ✅

Step 3:
What is the woman's location at the beginning of the video?
Prediction :
The woman is inside a house ✅

Step 4:
What is the woman's location at the end of the video?
Prediction :
The woman is outside, near the truck ✅

Omit similar content below

✅ **Final Prediction**
D:returning

✅ **Final Prediction**
D:returning

Figure 10: An example from the 150 seed samples. The top section shows the video input and corresponding counterfactual questions. The middle section compares two reasoning frameworks: the test-time long reasoning model QVQ and InternVL2.5-78B with CoT, with green marks indicating validated response components. The bottom section displays final model predictions, where green checkmarks indicate correct answers.

Input

Manual annotation

Video

Question
Is hillary clinton is giving a speech at the beginning of the video, if the video is played in reverse?
A: yes   B: no

Golden
B: no

Sub q1
What is hillary clinton doing at the beginning of the video, if the video is played in reverse?

Sub q2
What is hillary clinton doing at the end of the video, if the video is played in reverse?

QVQ

Alright, I've got this video to analyze..
It seems like there are a few different scenes here, and I need to think about the sequence of events if the video is reversed. Let me try to map out the scenes in the order they appear: 1. Large crowd with person at podium giving a speech. 2. Group of people (possibly officials or security) walking outdoors. 3. Hillary Clinton posing with children.\n\nIf the video is played in reverse, the order would be: **1. Hillary Clinton posing with children**. 2. Group of people walking outdoors. 3. **Large crowd with person at podium giving a speech**. So, in the reversed version, the speech comes at the end, not the beginning. But the question is specifically about the beginning of the reversed video. So, is Hillary Clinton giving a speech at the start of the reversed video? Based on my understanding, no, because in the reversed video, the first scene would be her posing with children, followed by the outdoor walking group, and then the speech. **Therefore, the answer should be B: no.**
Wait a minute, but I need to make sure I'm not missing something.
Let's see. The large crowd and podium suggest a political rally or event, possibly related to an election or campaign. The group of people walking outdoors could be part of the same event, perhaps arriving or departing. And the photo op with Hillary Clinton and children might be part of the event's agenda, focusing on family or education issues. If the video is reversed, the sequence **starts with the photo op**, then the outdoor walking, **then the speech**. So, again, the speech is not at the beginning in the reversed version. Perhaps the confusion comes from the fact that Hillary Clinton is known for giving speeches, especially in political contexts. But based on the images provided, in the reversed video, her speech would come after the other scenes, not before. **Therefore, I think the answer is B: no,** Hillary Clinton is not giving a speech at the beginning of the reversed video.
Just to double-check
Let's consider if there's any possibility that the speech and the photo op are simultaneous, but that seems unlikely. Typically, such events have a schedule: there might be speeches, then photo ops, or vice versa, but not both happening at the same time. So, unless there's something unusual about this particular event, I think my reasoning holds. Alright, I think I've thought this through sufficiently. **The answer is B: no.**

InternVL2.5-CoT

Step 1:
What is Hillary Clinton doing at the beginning of the video?
Prediction :
She is giving a speech to a large crowd. ✅

Step 2:
What is the context of the video?
Prediction :
The video shows Hillary Clinton giving a speech and interacting with people, including children ✅

Step 3:
Is the video played in reverse?
Prediction :
yes ✅

✅ **Final Prediction**
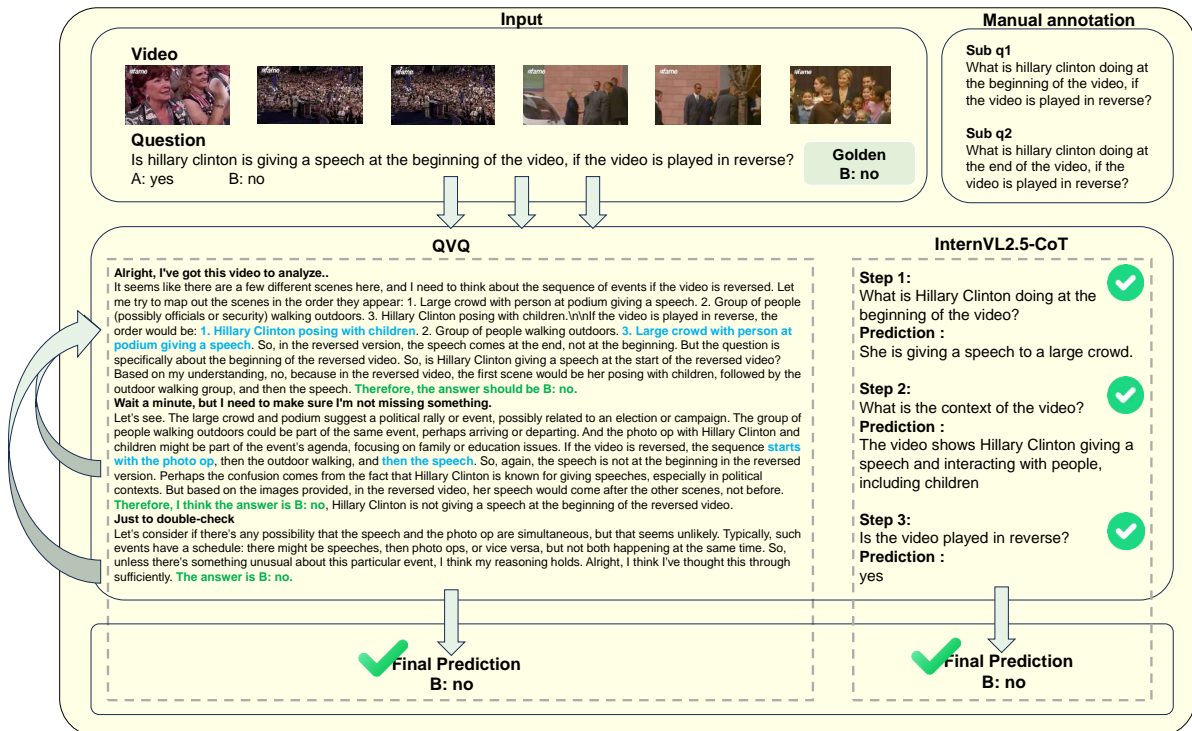B: no

✅ **Final Prediction**
B: no

Figure 11: An example from the 150 seed samples. The top section presents the video input and corresponding counterfactual questions. The middle section compares QVQ and InternVL2.5-78B with CoT, using a dual-color annotation scheme: blue indicates conceptual alignment with manual sub-problem annotations, and green highlights validated response components. The bottom section shows the final model predictions, with green checkmarks indicating correct answers.