V-STaR: Benchmarking Video-LLMs on Video Spatio-Temporal Reasoning

Zixu Cheng[1], Jian Hu[1]*, Ziquan Liu[1], Chenyang Si[2], Wei Li[3], Shaogang Gong[1]
[1]Queen Mary University of London, [2]Nanjing University, [3]Nanyang Technological University
{zixu.cheng,jian.hu,ziquan.liu,s.gong}@qmul.ac.uk,chenyang.si@nju.edu.cn,wei.l@ntu.edu.sg
https://V-STaR-Bench.github.io/

Abstract

*Human reasoning about video content typically follows a sequential process: establishing a spatio-temporal ("when-where") context before inferring "what" occurs. In contrast, current Video Large Language Models (Video-LLMs) often reverse this order, performing well on what-based video question answering while struggling with the preceding when and where grounding steps. Existing benchmarks mostly focus on "what" object or event presence, overlooking deeper relational and spatio-temporal reasoning. To address this gap, we propose a **V**ideo **S**patio-**T**empor**a**l **R**easoning (V-STaR) benchmark, designed to evaluate how Video-LLMs integrate explicit temporal and spatial cues in reasoning. The benchmark is built upon a Reverse Spatio-Temporal Reasoning (RSTR) mechanism that enforces a human-like Chain-of-Thought by decomposing video understanding into a coarse-to-fine series of interrelated questions about the same event. To support this evaluation, we construct a dataset using a semi-automated GPT-4-based pipeline that generates CoT-style spatio-temporal reasoning questions. Comprehensive evaluations of 16 state-of-the-art Video-LLMs on V-STaR reveal significant gaps in their ability to reason about "when-where" to "what" events occur. Our analysis uncovers three key insights: (1) models often rely on static representations instead of understanding dynamic processes; (2) there is a gap between implicit knowledge and explicit reasoning; and (3) semantic biases dominate the models' predictions. These findings highlight current limitations and point toward the need for more robust and interpretable video reasoning models.*

1. Introduction

Human video comprehension is a flexible sequential spatio-temporal reasoning process. As illustrated in Fig. 1, cognitive studies [18, 32] suggest that humans tend to reason by first constructing a coherent understanding of where and when events occur before inferring what happens. In other words, a stable spatio–temporal context serves as the foundation for higher-level semantic reasoning. This structured "where/when→what" process aligns with human intuition, yet current multimodal models often exhibit the opposite pattern: they can describe what occurs reasonably well, but struggle to localise where and when it happens. Such behaviour contradicts the expected reasoning order and raises a fundamental question: how do models actually understand spatio-temporal relations?

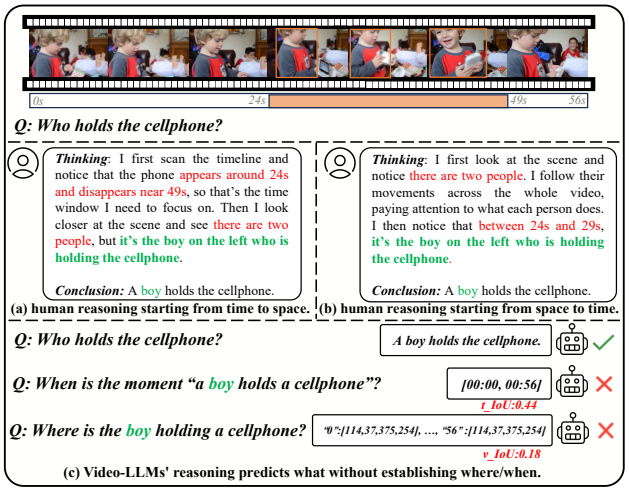


Figure 1. Difference between human flexible spatio-temporal reasoning and Video-LLM Limitations. (a) and (b) illustrate that human reasoning flexibly establishes "When/Where" context understanding as a necessary foundation to infer "What". In contrast, (c) shows that current Video-LLMs often correctly predict "What" without establishing the spatio-temporal reasoning "When/Where" steps. This motivates our benchmark's design to explicitly evaluate this gap and to analyze how models integrate explicit spatio-temporal cues in reasoning.

While this discrepancy highlights a fundamental gap between human and model reasoning, current video benchmarks seldom examine how models build spatio-temporal understanding. As shown in Tab. 1, most benchmarks focus on isolated aspects of what, when, or where, without evaluating their interdependence. This fragmented design obscures how models actually integrate spatial and tempo-

*corresponding author

Benchmark	VQA with Grounding			CoT Questions	Tasks
	VQA	Temporal	Spatial		
MVBench [20]	✓	-	-	-	MCQ
VideoMME [8]	✓	-	-	-	MCQ
LongVideoBench [35]	✓	-	-	-	MCQ
HourVideo [3]	✓	-	-	-	MCQ
MMBench-Video [6]	✓	-	-	-	Open-ended
QAEgo4D [2]	✓	✓	-	-	Open-ended
NeXT-GQA [36]	✓	✓	-	-	Open-ended
REXTIME [4]	✓	✓	-	-	Open-ended
E.T. Bench [24]	✓	✓	-	-	Open-ended
GCG [27]	✓	-	✓	-	Open-ended
Perception Test [30]	✓	✓	✓	-	Open-ended
VideoVista [21]	✓	✓	✓	-	MCQ
TVQA+ [17]	✓	✓	✓	-	Open-ended
V-STaR (Ours)	✓	✓	✓	✓	Open-ended

Table 1. Comparison of spatio-temporal understanding benchmarks. "VQA with Grounding" columns indicate whether each dataset involves visual question answering with temporal or spatial grounding requirements as spatio-temporal reasoning process.

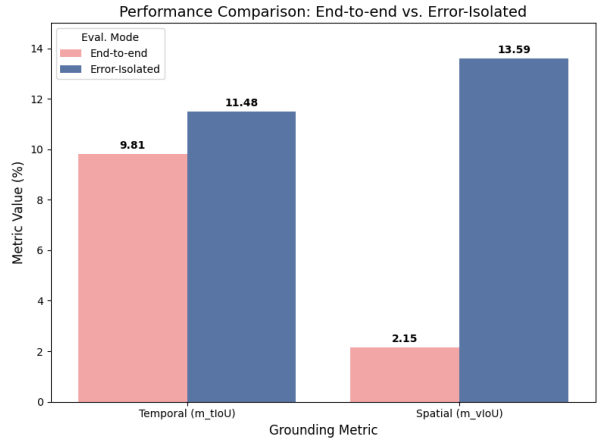


Figure 2. Comparison of end-to-end and error-isolated evaluation. Error-isolated evaluation provides a clearer measure of each reasoning step by preventing early errors from propagating, revealing latent spatio-temporal understanding hidden in end-to-end results.

ral cues, leaving unclear whether Video-LLMs genuinely ground their reasoning in coherent spatio-temporal relations or rely on superficial correlations. To address this gap, we propose a new benchmark that jointly evaluates video spatio-temporal reasoning, offering deeper insight into how Video-LLMs perceive spatio-temporal context as the foundation for reasoning.

In this work, we introduce **V**ideo **S**patio-**T**empor**a**l **R**easoning (V-STaR), a benchmark for explicitly evaluating the spatio-temporal reasoning ability of Video-LLMs. V-STaR has two key components. First, the Reverse Spatio-Temporal Reasoning (RSTR) task systematically decomposes a model's reasoning process. Unlike how human reasoning, RSTR uses reverse CoT paths ("what–when–where" and "what–where–when"), starting from the model's strength in what-based VQA and tracing back temporal and spatial evidence. Specifically, as shown in Fig. 2, end-to-end evaluation suffers from severe error accumulation. Early mistakes cascade through the chain. As a result, later failures become hard to interpret, since it is unclear whether they reflect weak reasoning or merely propagated errors. To address this, we employ an error-isolated evaluation, where each step receives ground-truth inputs from the previous one, preventing error propagation and enabling a clearer assessment of the model's intrinsic spatio-temporal reasoning. Second, we construct a fine-grained reasoning dataset using a semi-automated GPT-4-powered pipeline that generates structured CoT tasks for both RSTR paths. Each reasoning chain explicitly mimics human cognitive logic and is annotated with custom tags (e.g., `<think></think>`) to guide the model's reasoning process. Finally, we propose the Logarithmic Geometric Mean (LGM) metric to provide a unified assessment across all reasoning steps. Comprehensive experiments on 16 state-of-the-art Video-LLMs reveal that their reasoning relies on static cues, translates poorly into explicit spatio-temporal logic, and is strongly influenced by semantic bias, often resulting in plausible yet weakly grounded predictions. Our contributions are summarized as follows:

- We introduce V-STaR, the first benchmark explicitly designed to evaluate the human-like spatio-temporal reasoning ability of state-of-the-art Video-LLMs.
- We propose the Reverse Spatio-Temporal Reasoning (RSTR) task with a step-wise error-isolated evaluation and a Logarithmic Geometric Mean (LGM) metric, enabling systematic and interpretable assessment across coarse-to-fine reasoning steps.
- Extensive experiments on 16 Video-LLMs reveal that they rely on static cues, struggle to convert implicit understanding into explicit spatio-temporal reasoning, and are strongly influenced by semantic bias, often producing plausible yet weakly grounded predictions.

2. Related Works

Spatio-temporal understanding in Video-LLMs. Video-LLMs [11, 13, 19, 31, 34, 39, 43] have made rapid progress in video understanding, enabling them to answer a diverse range of questions about videos, e.g. framed as video question answering (VQA) problems. Many open-source Video-LLMs demonstrate competitive results to the proprietary commercial models, e.g., GPT-4o [29] and Gemini-2-Flash [10], across multiple Video-LLM Benchmarks [6, 8, 35]. Recent studies have explored the ability of Video-LLMs in video temporal and spatial understanding. TimeChat [31], VTimeLLM [31], and Trace [11] were among the first to develop specialized models for video temporal grounding, which involves localizing event timestamps in a video given a text description. Additionally, general-purpose models, such as Qwen2.5-VL [1] and VideoLlama3 [40], also exhibit strong temporal grounding capability in video, achieving comparable performance of

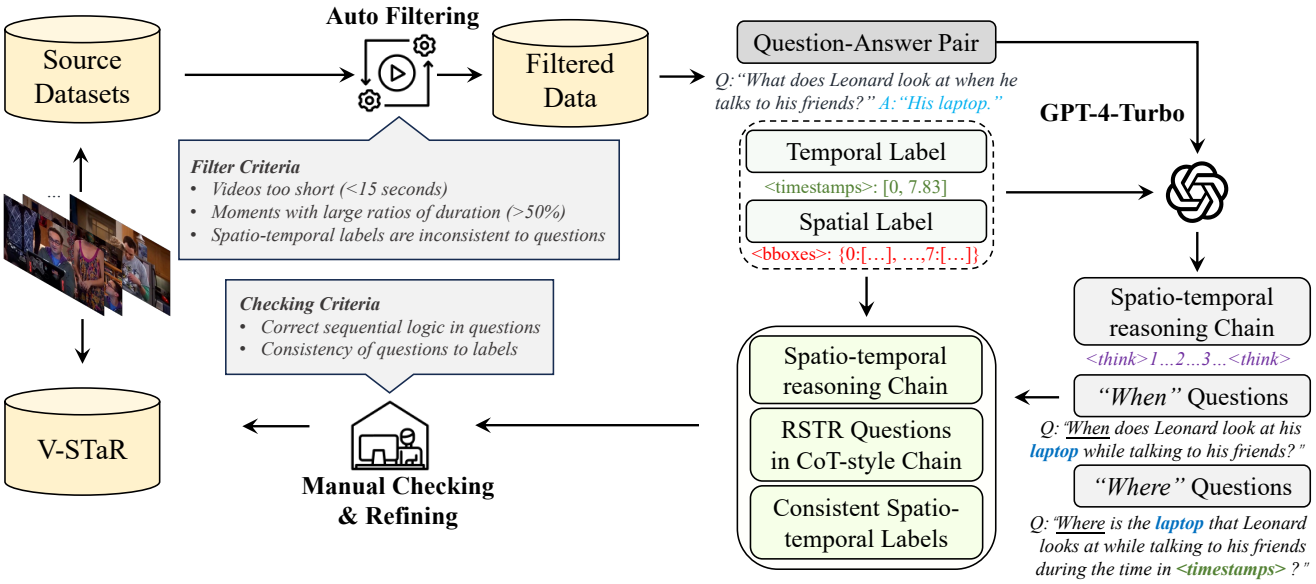


Figure 3. A semi-automated data construction pipeline of V-STaR: GPT-4-Turbo generates a spatio-temporal reasoning CoT chain to answer VQA questions, along with a set of RSTR questions. The RSTR questions are error-isolated temporal or spatial grounding challenges, decomposed from the CoT reasoning chain, designed to evaluate the model's spatio-temporal reasoning capabilities. Labels are human-annotated from sources and only leveraged to generate CoT-style questions in GPT-4-Turbo.

classic models [41, 42] on temporal grounding datasets [9, 15]. While certain Video-LLMs [1, 34, 40] claim to support object detection [22] and referring expression comprehension [38] on image inputs, their video spatial grounding capabilities remain largely unexplored. [26] first introduces spatial grounding to Video-LLMs, later extended to video segmentation [27, 37, 39]. However, most existing Video-LLMs evaluate their performance on VQA, temporal grounding, and spatial grounding tasks separately, without validating their ability for spatio-temporal reasoning. It is unclear whether Video-LLMs correctly understand and use spatio-temporal information in video reasoning.

Video-LLM Benchmarks. Recently, numerous benchmarks have been proposed to evaluate the general video understanding and reasoning capabilities of Video-LLMs. These benchmarks span a diverse range of tasks [20, 21, 23], types [6, 8, 33] and durations [3, 35, 45]. However, they primarily focus on Video Question Answering (VQA), essentially addressing the *"what"* question in videos while overlooking whether a model correctly understands and leverages spatio-temporal context in their reasoning process. To bridge this gap, some studies [4, 27, 36] have begun incorporating temporal or spatial grounding to validate the reasoning pathways of Video-LLMs. TVQA [16] proposed Grounded Video Question Answering (GVQA), requiring models to answer not only multiple-choice questions but also temporal grounded evidence in TV series videos. Expanding upon GVQA, benchmarks such as QAEgo4D [2], Next-GQA [36], and ReXTime [4] have extended these tasks to ego-centric videos, real-world videos, and complex reasoning questions. Grounded Conversation Generation (GCG) [27] was designed to challenge models in reasoning and identifying specific objects for segmentation in videos. VidSTG [44] further integrated spatio-temporal grounding with interrogative queries to reason the referred object in videos. TVQA+ [17] then introduced spatio-temporal grounding for VQA, but treated it as three independent sub-tasks, without investigating how models utilize temporal and spatial relationships in their reasoning process. Building on these works, our benchmark introduces CoT reasoning and employs temporal and spatial grounding as a structured reasoning chain, aiming to explicitly investigate the spatio-temporal reasoning abilities of Video-LLMs, providing a more comprehensive evaluation framework for reliable and trustworthy Video-LLMs in the future.

3. V-STaR Benchmark

In this section, we first define the Reverse Spatio-Temporal Reasoning (RSTR) task for evaluating the spatio-temporal reasoning capabilities of Video-LLMs. Then, we introduce a semi-automatic pipeline using GPT-4 [28], to generate coarse-to-fine RSTR questions to construct the dataset.

3.1. Task Definition

Most existing benchmarks [3, 6, 8, 20, 35] require models to directly answer complex reasoning problems, often without revealing their underlying reasoning process. Due to pre-trained co-occurrence biases, models may rely on prior knowledge rather than truly reasoning over the video, leading to inconsistencies such as hallucinations [12] in

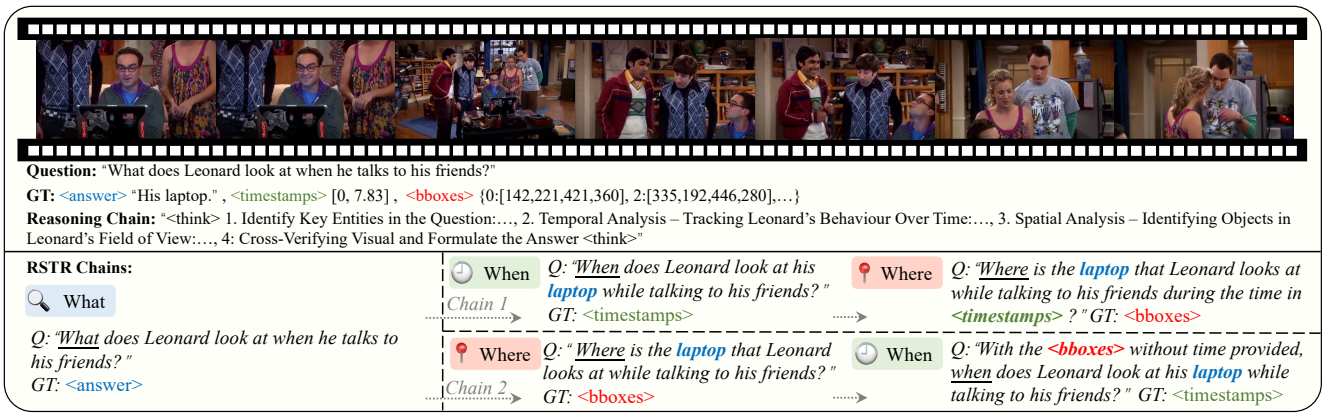


Figure 4. An example illustrating the construction of CoT questions. Each sample contains a thinking chain and two RSTR question chains: *"what-when-where"* and *"what-where-when"*.

the wrong time or place. To better assess the reasoning ability, we propose the Reverse Spatio-Temporal Reasoning (RSTR) task, based on three fundamental elements: *"what"*, *"when"*, and *"where"*. RSTR is inspired by how humans typically reason [32]: beginning with identifying relevant moments (*"when"*), then locating the key objects and their interactions (*"where"*), and finally answering the *"what"* question. To evaluate this, we adopt a Reverse Coarse-to-Fine CoT strategy: the model is first prompted to answer the *"what"* question, and then, based on that answer, a coarse-to-fine reasoning chain following the order *"what-when-where"* evaluates the model's spatio-temporal reasoning capability in a coarse-to-fine manner. Given the observation in Fig. 1, we also design a parallel chain in the order *"what-where-when"* to examine how different logical sequences impact the final results. Our RSTR task not only evaluates the model's spatio-temporal reasoning ability, but also quantifies the influence of various logical sequences.

3.2. Dataset Construction

A challenge in constructing this new dataset is to obtain videos with precise, coarse-to-fine CoT questions. To ease the burden of manual annotation, we exploited a hybrid approach that adapts annotated data from existing datasets while incorporating a semi-automated annotation pipeline. This approach consists of three stages: data collection, pipeline construction, and metric design.

Data Collection. We collected videos from datasets that offer spatial and temporal grounding. We used VidSTG [44], TVQA+ [17], and GOT-10K [14] datasets. VidSTG provides spatio-temporal grounding. TVQA+ offers temporal grounding for certain objects through question-answer pairs. GOT-10k gives spatial grounding details. However, these datasets do not include CoT reasoning chains, and their video durations are mostly in 0-3 minutes. Such rather short video durations are much narrower than what is seen in real-world scenarios. To ensure a diverse range of video durations, we started with the GOT-10k dataset because it has complete spatial grounding information. We then collected additional videos from YouTube that range from 3 minutes to 1 hour. Selected GOT videos were randomly inserted into various points within these videos. It ensures the final dataset is diverse in both duration and content, with the temporal labels marking where the GOT videos were inserted.

Pipeline Construction. While we collected a diverse set of videos with complete spatio-temporal labels, we aim to evaluate the model's spatio-temporal reasoning ability in a fine-grained manner. To achieve this, we leveraged GPT-4-Turbo [28] to construct a semi-automated pipeline for generating CoT reasoning chains and questions with a coarse-to-fine granularity. Specifically, as shown in Fig. 3, we first automatically filter out samples with short videos or overly large moment-to-video ratios, ensuring the questions remain sufficiently challenging. Next, we input the video question, answer, and corresponding temporal and spatial annotations into GPT-4-Turbo to generate a spatio-temporal reasoning chain for the question. The chain is then decomposed into two error-isolated fine-grained sub-questions focusing on temporal and spatial localization to evaluate whether the model's spatial and temporal reasoning is correct. Finally, two expert annotators manually review and refine the generated questions to ensure logical consistency and alignment with the video content. Importantly, all labels are human-annotated; GPT-4-Turbo is only used to generate questions with valid CoT-style and reasoning chains. The labels in the chains are directly adopted or transformed to fit our tasks.

Furthermore, to comprehensively investigate how a model leverages temporal and spatial cues during reasoning, we formulate the generated questions into two RSTR task chains: *"what-when-where"* and *"what-where-when"*. In each reasoning chain, the subsequent question incorporates the ground truth of the previous question. For in-

stance, in the *"what-when-where"* chain, the *"when"* question contains the ground truth of the *"what"* question, and the "where" question includes the ground truths of both the *"when"* and *"what"* questions. This design prevents the model from making errors in earlier reasoning steps and propagating to the final result, allowing for an error-isolated and fairer evaluation of temporal and spatial reasoning. Ultimately, each sample is associated with one spatio-temporal CoT reasoning chain and two RSTR task chains.

Metric Design. To evaluate the model's spatio-temporal reasoning ability, we have decomposed the task into fine-grained CoT reasoning questions, each targeting the *"what"*, *"when"*, or *"where"* aspect. These are individually assessed using accuracy (*Acc*), mean temporal IoU (m_tIoU), and mean visual IoU(m_vIoU). While this setup measures performance on each aspect, it overlooks the logical connections between them. Therefore, a unified metric is needed to better reflect the model's overall spatio-temporal reasoning ability. To overcome this problem, we propose evaluating the model's overall performance across these three questions using the Arithmetic Mean (AM) (Eq. 1) and a modified logarithmic Geometric Mean (LGM) (Eq. 3). Specifically, AM is given as:

$$AM = \frac{1}{3}(Acc + m\_tIoU + m\_vIoU), \tag{1}$$

while AM effectively assesses the model's overall performance across different metrics, it is susceptible to extreme values. To mitigate this issue, we employ the Geometric Mean (GM) to evaluate model performance:

$$GM = (Acc \times m\_tIoU \times m\_vIoU)^{\frac{1}{3}}, \tag{2}$$

However, when any of the metrics is zero, GM will become zero, which fails to reflect the contribution of the remaining metrics. To alleviate it, we transform GM into a logarithmic GM (LGM) as follows:

$$\begin{aligned} LGM = -\frac{1}{3}\Big\{ &\ln\left(1 - Acc + \epsilon\right) + \ln\left(1 - m\_tIoU + \epsilon\right) \\ &+ \ln\left(1 - m\_vIoU + \epsilon\right)\Big\}, \end{aligned} \tag{3}$$

where ϵ is a small constant to prevent $\ln(0)$ when any metric reaches 1. Eq.3 maps the metric range from 0 to positive infinity and ensures higher performance corresponds to a higher LGM score. Since the logarithm transformation results in values that are typically small in magnitude, we multiply LGM by a linear scaling factor of 100 to ensure numerical clarity, allowing finer distinctions between different methods while preserving relative ranking.

Moreover, when the same questions appear in different CoT chains, the order in which they occur can lead to significant variations in the results. To assess the overall performance of the model across different chains, we propose the mean AM (mAM) and mean LGM (mLGM) as follows:

$$mAM = \frac{1}{n}\sum_{k=1}^{n} AM_k, \quad mLGM = \frac{1}{n}\sum_{k=1}^{n} LGM_k. \tag{4}$$

where n denotes the number of different chains. The mAM and mLGM effectively evaluate the combined impact of the various chains on the model's performance.

3.3. Dataset Statistics

Here, we present detailed statistics of our dataset, including video information, meta information, qualitative analyses, and comparisons with previous works.

Video Information. Our dataset comprises 2094 videos totalling 64.12 hours of footage. As shown in Fig.5(a), to ensure the inclusion of varied video genres, we categorized the videos into 9 domains: Entertainment, Daily Life, Indoor, Sports, Animals, Vehicles, Nature, Shows, and Tutorial. The length distribution of the videos, illustrated in Fig.5(b), demonstrates considerable diversity. The videos range in length from 15.02 seconds to 59.2 minutes with average 110.23 seconds, satisfying the requirement for diverse video lengths and better reflecting real-world scenarios.

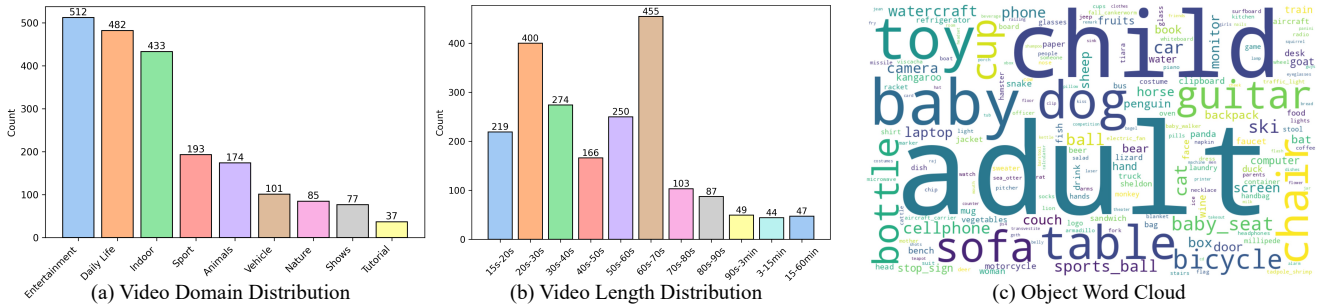
Meta Information. To evaluate dataset completeness, we analyzed the meta-information annotations. Each video is accompanied by temporal moment annotations, with an average duration of 9.06 seconds and individual durations ranging from 1.7 seconds to 47 seconds. These temporal moments account for an average of 19.3% of the total video duration, ensuring a reasonable level of difficulty for the temporal grounding subtask. For the spatial grounding subtask, we annotated 342 objects with a total of 16,793 bounding boxes, covering approximately 19.8% of the video resolution. This proportion is similar to that of the temporal grounding, ensuring consistent challenge levels across both tasks. Additionally, we visualized the object categories with a word cloud (Fig.5(c)), demonstrating that our questions robustly capture a wide diversity of objects. Tab.2 provides further detailed statistics.

Qualitative analyses. Fig. 4 shows an example from our V-STaR benchmark. It contains one spatio-temporal CoT reasoning thinking chain and two RSTR task chains. For each RSTR task chain, the CoT evaluation starts with a coarse-grained question about *"what"* in the video. In the *"what-when-where"* chain, the subsequent *"when"* question incorporates the answer of ‘*"what"* and its answer is included in the *"where"* question. In the other chain, the subsequent *"where"* question contains the answer of *"what"* and the bounding boxes answer of *"where"* will be provided without time information in the *"when"* question.

Comparisons with previous benchmarks. Tab. 1 compares V-STaR with existing benchmarks. Most existing

	Entertainment	Daily Life	Indoor	Sports	Animals	Vehicles	Nature	Shows	Tutorial	Overall
Avg Length(s)	104.60	88.21	45.24	128.00	38.07	42.16	44.19	258.14	1512.05	110.23
Avg Moment(s)	9.32	8.68	10.40	6.99	8.10	7.70	8.96	10.71	10.45	9.06
Avg M/L Ratio(%)	15.16	20.29	22.98	20.76	21.30	20.01	20.49	18.34	2.02	19.32
Num of BBox	2097	4351	4621	1409	1471	806	789	840	409	16793
Num of Objects	255	38	29	37	26	16	12	18	29	342

Table 2. Statistical comparison of different domains.



(a) Video Domain Distribution (b) Video Length Distribution (c) Object Word Cloud

Figure 5. Dataset statistics of video domain and length, and visualization of objects in video.

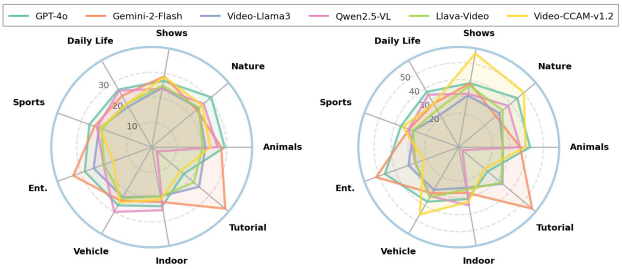
datasets focus only on *"what"* question in VQA [3, 6, 8, 20, 23, 33, 35], lacking evaluation of spatio-temporal reasoning. Some partially cover on *"when"* [2, 4, 16, 24, 36] or *"where"* [27], without complete spatio-temporal reasoning chain. [17, 21, 30] covered all, but they ignored their inner spatio-temporal reasoning relationship. Instead, our V-STaR provides two CoT question chains for each sample to reveal the spatio-temporal reasoning ability of Video-LLMs.

4. Experiments

4.1. Setting and Metrics

Implementation Details. We tested 16 Video-LLMs, involving 2 commercial models GPT-4o [29] and Gemini-2-Flash [10], and 14 open-source models. The open-source models include (i) 10 generic models: Video-LLaMA3 [40], Qwen2.5-VL(7B/32B) [1], Qwen2-VL [34], InternVL-2.5(7B/32B) [5], LLaVA-Video [43], VideoChat2 [20], Oryx-1.5 [25], and Video-CCAM-v1.2 [7]); (ii) 3 time-aware models: TimeChat [31], VTimeLLM [13], and Trace [11]; and (iii) 1 segmentation model, Sa2VA [39]. We followed their official configurations and sampled the video frames at 1fps for all models. If a video exceeded the model's input limitations, we applied uniform sampling to select the maximum allowable number of frames. We investigated the models' spatio-temporal reasoning ability using two RSTR task chains: *"what-when-where"* and *"what-where-when"*. Experiments were run on 4 NVIDIA A100 80G GPUs.

Metrics. To evaluate the open-ended *"what"* question, we follow MMBench-Video [6] and use Qwen2.5-72B [1] to score answers from 0 to 3, denoting *"entirely incorrect"*, *"largely incorrect"*, *"largely correct"*, and *"entirely correct"*. Answers scoring above 2 are considered correct, allowing us to compute accuracy. For the *"when"* question,



(a) mAM across domains (b) mLGM across domains

Figure 6. The performance of each domain.

we follow the commonly used temporal grounding metrics, "R@n, tIoU=m", which refers to the percentage of top-n prediction with temporal IoU score larger than m, and mean temporal IoU score (m_tIoU). For the *"where"* question, we follow TVQA+ [17] and VidSTG [44] to use the Average Precision score (AP@vIoU=m) and mean visual Intersection over Union (m_vIoU) of every annotated frame. We follow the proposed LGM (Eq.3) and AM (Eq.1) to measure a model's spatial-temporal reasoning ability. A higher LGM indicates a better overall spatio-temporal reasoning ability of the model, and a higher AM indicates a more average performance of the model on the three metrics.

4.2. Quantitative Results

We evaluate Video-LLMs under two reasoning chains, *"what-when-where"* and *"what-where-when"*, as summarised in Tab. 3. Each chain measures the model's ability to infer event semantics (*what*), temporal localisation (*when*), and spatial grounding (*where*) in a step-wise manner. To provide a comprehensive understanding, we further analyse domain-wise performance (Fig. 6), the effect of video length (Tab. 4), and joint reasoning consistency across chains (Tab. 5). The quantitative and qualitative results collectively reveal three central insights into how Video-LLMs reason about space and time: (i) Static representation sub-

Model	Params	What (VQA)		Chain 1: What-When-Where						Chain 2: What-Where-When					
				Temporal		Spatial		LGM	AM	Spatial		Temporal		LGM	AM
		Score	Acc	R1@0.5	m_tIoU	AP@0.5	m_vIoU			AP@0.5	m_vIoU	R1@0.5	m_tIoU		
GPT-4o [29]	-	1.71	**60.78**	10.35	16.67	2.75	6.47	**39.51**	**27.97**	1.19	3.01	10.04	12.82	**36.79**	25.53
Gemini-2-Flash [10]	-	1.59	53.01	15.84	**24.54**	0.93	4.63	36.14	27.39	0.58	2.21	15.22	**23.83**	34.99	**26.35**
Video-LLaMA3 [40]	7B	1.38	41.94	**19.80**	22.97	0.11	0.89	27.12	21.93	0.02	0.19	**20.42**	23.14	26.96	21.76
Qwen2.5-VL [1]	7B	1.61	54.53	8.92	11.48	8.36	13.59	35.20	26.53	1.40	2.00	5.39	7.61	29.58	21.38
Qwen2.5-VL [1]	32B	1.65	59.26	7.07	9.90	7.63	11.91	37.64	27.03	2.58	3.51	4.01	6.28	33.29	23.02
Qwen2-VL [34]	7B	1.03	25.91	17.94	19.18	3.89	9.31	20.35	18.13	1.14	2.41	16.32	17.52	17.23	15.28
InternVL-2.5 [5]	8B	1.46	44.18	4.87	8.72	0.04	0.65	22.69	17.85	0.00	0.14	3.77	7.75	27.15	17.36
InternVL-2.5 [5]	38B	1.48	48.14	8.26	15.67	1.73	6.22	29.71	23.34	0.32	1.34	9.41	17.49	28.74	22.32
Llava-Video [43]	7B	1.50	49.48	6.30	10.52	0.18	1.92	27.11	20.64	0.25	1.31	5.49	12.21	27.54	21.00
VideoChat2 [20]	7B	1.27	36.21	13.07	13.69	0.14	2.51	20.74	17.47	0.30	0.97	12.07	12.50	19.77	16.56
Oryx-1.5 [25]	7B	0.94	20.47	4.48	13.54	2.17	10.14	16.05	14.72	0.96	3.50	5.58	14.81	14.16	12.93
Video-CCAM-v1.2 [7]	7B	**1.75**	59.35	0.00	1.50	-	-	30.51	20.28	-	-	0.00	2.26	30.88	20.54
TimeChat [31]	7B	1.06	26.38	8.68	12.01	-	-	14.47	12.80	-	-	8.54	13.60	15.08	13.33
VTimeLLM [13]	7B	1.45	41.46	10.88	17.13	0.03	0.21	24.18	19.60	0.00	0.00	4.53	5.96	19.90	15.81
TRACE [11]	7B	0.90	17.60	14.17	19.74	-	-	13.78	12.45	-	-	12.02	17.11	12.71	11.57
Sa2VA [39]	8B	0.70	16.36	0.00	0.11	**34.18**	**32.31**	19.00	16.26	**40.42**	**37.48**	0.00	0.00	21.61	17.95

Table 3. Performance Comparison of Video-LLMs Across Two Reverse Spatio-Temporal Reasoning (RSTR) Chains. The top result is highlighted in **bold**, while the second is underlined. "-" denotes a model that failed to generate formatted answers. The score ranges from 0 to 3, showing the quality of the model's open-ended answer.

Model	Short		Medium		Long		All	
	mAM	mLGM	mAM	mLGM	mAM	mLGM	mAM	mLGM
GPT-4o [29]	**27.49**	**38.56**	26.96	**40.58**	14.86	19.28	26.75	**38.15**
Gemini-2-Flash [10]	24.97	32.07	**28.99**	40.35	**37.81**	**56.14**	**26.87**	35.57
Video-LLaMA3 [40]	21.68	26.62	21.84	27.23	22.46	28.83	21.66	27.04
Qwen2.5-VL [1]	25.51	34.84	23.67	32.87	2.20	2.27	23.96	32.39
Qwen2-VL [34]	15.78	17.50	18.47	21.22	14.09	17.53	16.71	18.79
InternVL-2.5 [5]	17.94	22.90	17.94	23.06	9.58	11.19	17.60	24.92
Llava-Video [43]	22.37	30.23	18.28	22.77	18.23	25.23	20.82	27.33
VideoChat2 [20]	17.57	21.02	17.20	20.50	5.28	5.64	17.02	20.26
Oryx-1.5 [25]	13.17	14.25	14.83	16.46	11.89	13.99	15.11	13.83
Video-CCAM-v1.2 [7]	21.66	34.09	19.62	28.36	12.61	15.80	20.41	30.70
TimeChat [31]	13.70	15.56	13.22	15.06	3.24	3.37	13.07	14.78
VTimeLLM [13]	18.31	23.19	18.15	22.44	5.52	5.89	17.71	22.04
TRACE [11]	11.77	12.96	12.49	13.87	13.59	15.30	12.01	13.25
Sa2VA [39]	18.14	22.01	16.32	18.92	8.85	9.70	17.11	20.31

Table 4. Performance on different video lengths. "Short", "Medium" and "Long" denote durations of $[0, 1]$ min, $(1, 3]$ min, and $(3, 60]$ min, respectively. The top result is in **bold**, and the second is underlined.

stitutes dynamic understanding, (ii) Implicit understanding vs. explicit reasoning gap, and (iii) Semantic bias dominates reasoning.

(1) Static representation substitutes dynamic understanding. Tab. 3 presents the results of the *"what-when-where"* chain, where models sequentially answer semantic, temporal, and spatial questions. While GPT-4o, Gemini-2-Flash, and Qwen2.5-VL-32B achieve the highest overall scores, all models exhibit a sharp decline from "what" prediction to temporal and spatial grounding, with m_tIoU and m_vIoU remaining below 20%. This suggests that current Video-LLMs depend largely on static appearance matching rather than dynamic motion reasoning. In open-source models, Video-LLaMA3-7B performs most consistently, while Qwen2.5-VL-7B balances moderate performance across all three sub-tasks. Moreover, as shown in Tab. 4, reasoning performance further deteriorates as video duration increases: GPT-4o performs well on short clips but drops notably on long sequences, whereas Gemini-2-Flash remains more stable, implying hierarchical temporal encoding benefits. Together, these results indicate that most Video-LLMs behave as *frame-wise captioners*, correlating isolated visual tokens with text rather than modelling causal transitions or event dynamics. This phenomenon is visually confirmed in Fig. 7, where models often treat moving objects as static and fail to maintain motion continuity.

(2) Implicit understanding vs. explicit reasoning gap. The *"what-where-when"* chain in Tab. 3 reverses the reasoning order to test whether models can explicitly compose temporal cues after spatial grounding. Most models experience substantial performance degradation—especially Qwen2.5-VL-7B, whose spatial IoU drops from 13.6% to 2.0%. This order sensitivity reveals that Video-LLMs possess certain latent spatio-temporal knowledge but lack a structured reasoning mechanism to propagate evidence between steps. Tab. 5 further supports this finding: when provided with intermediate ground-truth cues, spatial accuracy increases markedly (2.15%→13.59%), showing that models implicitly encode relational cues but fail to externalise them into explicit reasoning chains. The low joint accuracy for spatio-temporal reasoning (e.g., 4.68% for Qwen2.5-VL-7B in Chain 1 and 2.24% for Gemini-2-Flash in Chain 2) underscores this limitation. Hence, while Video-LLMs contain rich multimodal embeddings, they still lack explicit reasoning pathways comparable to human spatio-temporal cognition.

(3) Semantic bias dominates reasoning. Across both chains, models display strong bias toward semantic priors. As shown in Tabs. 3, "what" accuracy remains high even when grounding fails, indicating heavy dependence on linguistic correlations rather than perceptual verification. For

Figure 7. An example showcasing the performance of five Video-LLMs.

Model	Acc|tIoU@0.3			Acc|vIoU@0.1			Acc|tIoU@0.3, vIoU@0.1		
	Chain 1	Chain 2	|Δ|	Chain 1	Chain 2	|Δ|	Chain 1	Chain 2	|Δ|
GPT-4o [29]	15.12	11.16	**3.96**	15.27	**7.59**	7.68	4.53	**3.91**	0.62
Gemini-2-Flash [10]	**19.70**	**19.04**	0.66	8.68	4.48	4.20	3.48	2.24	1.24
Video-LLaMA3 [40]	15.41	14.89	0.52	1.34	0.19	1.15	0.52	0.05	0.47
Qwen2.5-VL [1]	10.73	7.20	3.53	**24.24**	4.25	**19.99**	**4.68**	1.15	**3.53**
Qwen2-VL [34]	8.06	6.68	1.38	7.11	2.05	5.06	2.29	1.53	0.76
InternVL-2.5 [5]	5.92	4.77	1.15	0.81	0.19	0.62	0.19	0.05	0.14
Llava-Video [43]	7.92	8.97	1.05	3.05	2.86	0.19	0.67	0.86	0.19
VideoChat-2 [20]	8.78	7.73	1.05	3.34	1.24	2.10	0.29	0.62	0.33
Oryx-1.5 [25]	3.58	3.77	0.19	6.25	2.05	4.20	1.24	0.67	0.57

Table 5. Joint performance evaluation across models. "Chain 1" and "Chain 2" denote reasoning orders, while $|\Delta|$ measures the consistency gap between them.

example, contextual phrases such as "talks to his friends" or "on the grass" often trigger memorised associations ("laptop", "bat") that are semantically plausible but visually unsupported. Domain-wise results in Fig. 6 reinforce this trend: GPT-4o and Gemini-2-Flash excel in language-rich contexts (e.g., *Animals*, *Daily Life*), but all models degrade sharply in visually complex or low-prior domains like *Tutorials*. This shows that current Video-LLMs primarily rely on *semantic correlation* instead of *causal perception*, resulting in poor generalisation across unseen visual domains.

Qualitative analysis. Fig. 7 visually supports the above insights through qualitative comparisons of five representative models. Although all can correctly answer *"what"* questions, their spatio-temporal reasoning remains inconsistent. In the *"what-when-where"* chain, Qwen2.5-VL-7B accurately localises objects but misorders events temporally, whereas InternVL-2.5-8B shows the reverse pattern. GPT-4o, Gemini-2-Flash, and Video-LLaMA3 are relatively balanced but still prone to spatial drift or coarse temporal boundaries. In the reverse *"what-where-when"* chain, performance degrades once temporal cues are missing—except for Qwen2.5-VL-7B, which retains stable spatial grounding. Across all models, frame-wise processing dominates, causing them to overlook object dynamics and causal transitions over time. This static-frame perception visually corroborates the quantitative conclusion that Video-LLMs rely on correlation rather than coherent temporal reasoning.

5. Conclusion

We present V-STaR, a benchmark for evaluating the spatio-temporal reasoning ability of Video-LLMs. It introduces the Reverse Spatio-Temporal Reasoning (RSTR) task with step-wise error-isolated evaluation and a Logarithmic Geometric Mean (LGM) metric for fair assessment. Experiments on 16 Video-LLMs show that while models perform well at recognizing "what", they struggle to find "when/where" cue in the video. We highlight three key insights: (1) static representations substitute for dynamic understanding, (2) a gap exists between implicit knowledge and explicit reasoning, and (3) semantic bias dominates predictions. These findings call for structured CoT training with explicit spatio-temporal supervision toward more causal and interpretable Video-LLMs.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2, 3, 6, 7, 8

[2] Leonard Bärmann and Alex Waibel. Where did i leave my keys?-episodic-memory-based question answering on egocentric videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1560–1568, 2022. 2, 3, 6

[3] Keshigeyan Chandrasegaran, Agrim Gupta, Lea M Hadzic, Taran Kota, Jimming He, Cristóbal Eyzaguirre, Zane Durante, Manling Li, Jiajun Wu, and Fei-Fei Li. Hourvideo: 1-hour video-language understanding. *Advances in Neural Information Processing Systems*, 37:53168–53197, 2025. 2, 3, 6

[4] Jr-Jen Chen, Yu-Chien Liao, Hsi-Che Lin, Yu-Chu Yu, Yen-Chun Chen, and Yu-Chiang F Wang. Rextime: A benchmark suite for reasoning-across-time in videos. *Advances in Neural Information Processing Systems*, 37:28662–28673, 2024. 2, 3, 6

[5] Zhe Chen, Weiyun Wang, Yue Cao, Yangzhou Liu, Zhangwei Gao, Erfei Cui, Jinguo Zhu, Shenglong Ye, Hao Tian, Zhaoyang Liu, et al. Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. *arXiv preprint arXiv:2412.05271*, 2024. 6, 7, 8

[6] Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Advances in Neural Information Processing Systems*, 37:89098–89124, 2024. 2, 3, 6

[7] Jiajun Fei, Dian Li, Zhidong Deng, Zekun Wang, Gang Liu, and Hui Wang. Video-ccam: Enhancing video-language understanding with causal cross-attention masks for short and long videos. *arXiv preprint arXiv:2408.14023*, 2024. 6, 7

[8] 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. 2, 3, 6

[9] Jiyang Gao, Chen Sun, Zhenheng Yang, and Ram Nevatia. Tall: Temporal activity localization via language query. In *Proceedings of the IEEE international conference on computer vision*, pages 5267–5275, 2017. 3

[10] Google. Google, gemini-2-flash. Technical report, Google, 2024. 2, 6, 7, 8

[11] Yongxin Guo, Jingyu Liu, Mingda Li, Xiaoying Tang, Qingbin Liu, and Xi Chen. Trace: Temporal grounding video llm via causal event modeling. *arXiv preprint arXiv:2410.05643*, 2024. 2, 6, 7

[12] Jian Hu, Jiayi Lin, Junchi Yan, and Shaogang Gong. Leveraging hallucinations to reduce manual prompt dependency in promptable segmentation. *Advances in Neural Information Processing Systems*, 37:107171–107197, 2025. 3

[13] Bin Huang, Xin Wang, Hong Chen, Zihan Song, and Wenwu Zhu. Vtimellm: Empower llm to grasp video moments. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14271–14280, 2024. 2, 6, 7

[14] Lianghua Huang, Xin Zhao, and Kaiqi Huang. Got-10k: A large high-diversity benchmark for generic object tracking in the wild. *IEEE transactions on pattern analysis and machine intelligence*, 43(5):1562–1577, 2019. 4

[15] Ranjay Krishna, Kenji Hata, Frederic Ren, Li Fei-Fei, and Juan Carlos Niebles. Dense-captioning events in videos. In *Proceedings of the IEEE international conference on computer vision*, pages 706–715, 2017. 3

[16] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara Berg. Tvqa: Localized, compositional video question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1369–1379, 2018. 3, 6

[17] Jie Lei, Licheng Yu, Tamara Berg, and Mohit Bansal. Tvqa+: Spatio-temporal grounding for video question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8211–8225, 2020. 2, 3, 4, 6

[18] Jiangtong Li, Li Niu, and Liqing Zhang. From representation to reasoning: Towards both evidence and commonsense reasoning for video question-answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 21273–21282, 2022. 1

[19] KunChang Li, Yinan He, Yi Wang, Yizhuo Li, Wenhai Wang, Ping Luo, Yali Wang, Limin Wang, and Yu Qiao. Videochat: Chat-centric video understanding. *arXiv preprint arXiv:2305.06355*, 2023. 2

[20] 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. 2, 3, 6, 7, 8

[21] Yunxin Li, Xinyu Chen, Baotian Hu, Longyue Wang, Haoyuan Shi, and Min Zhang. Videovista: A versatile benchmark for video understanding and reasoning. *arXiv preprint arXiv:2406.11303*, 2024. 2, 3, 6

[22] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer vision–ECCV 2014: 13th European conference, zurich, Switzerland, September 6-12, 2014, proceedings, part v 13*, pages 740–755. Springer, 2014. 3

[23] Yuanxin Liu, Shicheng Li, Yi Liu, Yuxiang Wang, Shuhuai Ren, Lei Li, Sishuo Chen, Xu Sun, and Lu Hou. Tempcompass: Do video llms really understand videos? *arXiv preprint arXiv:2403.00476*, 2024. 3, 6

[24] Ye Liu, Zongyang Ma, Zhongang Qi, Yang Wu, Ying Shan, and Chang Wen Chen. Et bench: Towards open-ended event-level video-language understanding. *arXiv preprint arXiv:2409.18111*, 2024. 2, 6

[25] 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. 6, 7, 8

[26] Shehan Munasinghe, Rusiru Thushara, Muhammad Maaz, Hanoona Abdul Rasheed, Salman Khan, Mubarak Shah, and Fahad Khan. Pg-video-llava: Pixel grounding large video-language models. *arXiv preprint arXiv:2311.13435*, 2023. 3

[27] Shehan Munasinghe, Hanan Gani, Wenqi Zhu, Jiale Cao, Eric Xing, Fahad Shahbaz Khan, and Salman Khan. Videoglamm: A large multimodal model for pixel-level visual grounding in videos. *arXiv preprint arXiv:2411.04923*, 2024. 2, 3, 6

[28] OpenAI. Gpt-4 technical report. Technical report, OpenAI, 2023. 3, 4

[29] OpenAI. Gpt-4o system card. Technical report, OpenAI, 2024. 2, 6, 7, 8

[30] Viorica Patraucean, Lucas Smaira, Ankush Gupta, Adria Recasens, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Mateusz Malinowski, Yi Yang, Carl Doersch, et al. Perception test: A diagnostic benchmark for multimodal video models. *Advances in Neural Information Processing Systems*, 36:42748–42761, 2023. 2, 6

[31] Shuhuai Ren, Linli Yao, Shicheng Li, Xu Sun, and Lu Hou. Timechat: A time-sensitive multimodal large language model for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14313–14323, 2024. 2, 6, 7

[32] Camilo Miguel Signorelli, Selma Dündar-Coecke, Vincent Wang, and Bob Coecke. Cognitive structures of space-time. *Frontiers in Psychology*, 11:527114, 2020. 1, 4

[33] Enxin Song, Wenhao Chai, Guanhong Wang, Yucheng Zhang, Haoyang Zhou, Feiyang Wu, Haozhe Chi, Xun Guo, Tian Ye, Yanting Zhang, et al. Moviechat: From dense token to sparse memory for long video understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18221–18232, 2024. 3, 6

[34] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, et al. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *arXiv preprint arXiv:2409.12191*, 2024. 2, 3, 6, 7, 8

[35] Haoning Wu, Dongxu Li, Bei Chen, and Junnan Li. Longvideobench: A benchmark for long-context interleaved video-language understanding. *Advances in Neural Information Processing Systems*, 37:28828–28857, 2025. 2, 3, 6

[36] Junbin Xiao, Angela Yao, Yicong Li, and Tat-Seng Chua. Can i trust your answer? visually grounded video question answering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13204–13214, 2024. 2, 3, 6

[37] Cilin Yan, Haochen Wang, Shilin Yan, Xiaolong Jiang, Yao Hu, Guoliang Kang, Weidi Xie, and Efstratios Gavves. Visa: Reasoning video object segmentation via large language models. In *European Conference on Computer Vision*, pages 98–115. Springer, 2024. 3

[38] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. Modeling context in referring expressions. In *Computer Vision–ECCV 2016: 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part II 14*, pages 69–85. Springer, 2016. 3

[39] Haobo Yuan, Xiangtai Li, Tao Zhang, Zilong Huang, Shilin Xu, Shunping Ji, Yunhai Tong, Lu Qi, Jiashi Feng, and Ming-Hsuan Yang. Sa2va: Marrying sam2 with llava for dense grounded understanding of images and videos. *arXiv preprint arXiv:2501.04001*, 2025. 2, 3, 6, 7

[40] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, Yuqian Yuan, Guanzheng Chen, Sicong Leng, Yuming Jiang, Hang Zhang, Xin Li, et al. Videollama 3: Frontier multimodal foundation models for image and video understanding. *arXiv preprint arXiv:2501.13106*, 2025. 2, 3, 6, 7, 8

[41] Hao Zhang, Aixin Sun, Wei Jing, and Joey Tianyi Zhou. Span-based localizing network for natural language video localization. *arXiv preprint arXiv:2004.13931*, 2020. 3

[42] Songyang Zhang, Houwen Peng, Jianlong Fu, and Jiebo Luo. Learning 2d temporal adjacent networks for moment localization with natural language. In *Proceedings of the AAAI conference on artificial intelligence*, pages 12870–12877, 2020. 3

[43] 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. 2, 6, 7, 8

[44] Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10668–10677, 2020. 3, 4, 6

[45] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 3