Towards Temporal Compositional Reasoning in Long-Form Sports Videos

Siyu Cao[1,2], Lu Zhang[✉1,2], Ruizhe Zeng[1,2], and Zhi-yong Liu[✉1,2,3]

[1] MAIS, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China
{caosiyu2024, lu.zhang, zengruizhe2022, zhiyong.liu}@ia.ac.cn
[2] School of Artificial Intelligence, University of Chinese Academy of Sciences, Beijing 100049, China
[3] Nanjing Artificial Intelligence Research of IA, Nanjing, 211100, China

Abstract. Sports video analysis is a challenging domain for multimodal understanding because it involves complex and dynamic human activities. Despite rapid progress in Multimodal Large Language Models (MLLMs), long-horizon reasoning in sports videos remains difficult, as answering questions requires both locating and integrating temporally sparse evidence into reasoning. We attribute this limitation to two closely related factors: insufficient supervision over temporally dispersed evidence and the lack of methods for explicit temporal evidence localization and justification. To address these gaps, we introduce **SportsTime**, a large-scale benchmark for long-form sports video understanding, comprising 14K+ open-ended QA pairs and 50K+ step-wise temporal evidence annotations. Building on SportsTime, we propose **Chain-of-Time Reasoning (CoTR)**, which treats reasoning as a process of temporally grounded evidence composition. Specifically, during training, CoTR introduces a temporal-reward GRPO to encourage temporally grounded reasoning. During inference, it employs an anchor-observe-infer evidence-seeking loop to iteratively localize, verify, and compose temporal evidence before producing the final answer. Experiments show that SportsTime exposes substantial gaps in current MLLMs, while CoTR yields consistent gains over strong baselines, improving both temporal compositional reasoning performance and step-wise grounding quality. The dataset and code are available at `https://github.com/ustiniansy/SportsTime`.

Keywords: Temporal Compositional Reasoning · Sports · Benchmark

1 Introduction

Sports videos capture complex and dynamic human activities at scale, playing a central role in global cultural life and serving as a key data source for professional analytics. Over the past decade, artificial intelligence and computer vision techniques have fundamentally reshaped sports video analysis, enabling tasks such as

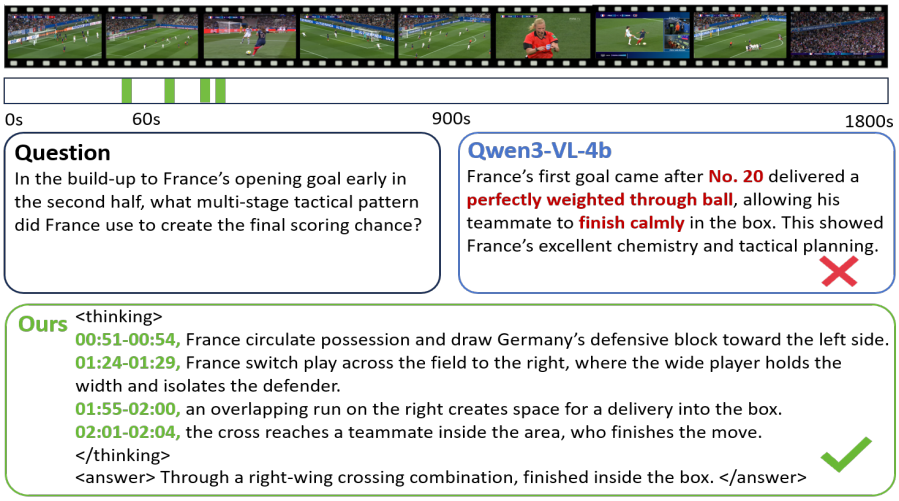


Fig. 1: Chain-of-Time reasoning enables more reliable and verifiable answers in long-form sports videos.

player detection and tracking, action spotting, tactical analysis, etc [7, 8, 12, 43]. More recently, the rapid development of Multimodal Large Language Models (MLLMs) [2, 6, 10] has opened new opportunities toward unified sports video understanding, particularly for more diverse and open-ended tasks such as commentary generation [27] and video question answering (VideoQA) [41], which require flexible and compositional inference over long-form multimodal content.

Despite significant advances, current MLLMs still struggle with long-horizon reasoning in videos [17, 18, 23, 28, 36, 54]. This weakness is particularly evident in sports scenarios, which are long-form and highly dynamic, where interpreting sparse but critical events demands a holistic and long-horizon understanding. For example, in a soccer match, understanding how a goal is scored may require tracing a sequence of events, such as passes and player movements that unfold long before the final shot. Such scenarios expose a key limitation of current MLLMs: they struggle to identify temporally dispersed evidence and compose it into reasoning, a capability we refer to as temporal compositional reasoning.

We argue that this limitation mainly stems from two tightly coupled aspects: (1) the scarcity of high-quality annotations that explicitly capture temporally dispersed evidence across multiple time spans, resulting in insufficient supervision for long-horizon reasoning [4, 34]; and (2) the lack of methods that explicitly encourage models to identify, localize, and justify the specific temporal evidence underlying their final answers [21]. As a result, models tend to overrely on language priors while underutilizing visual and temporal evidence, producing seemingly plausible yet weakly grounded answers and thus leading to hallucinations.

Motivated by this, we introduce **SportsTime**, a large-scale benchmark for comprehensive long-form sports video understanding through temporal composi-

tional reasoning. SportsTime comprises over 14,000 high-quality open-ended QA pairs across five representative team sports, spanning both highlight videos ($\sim$ 10 min) and full-game broadcasts ($\sim$ 50 min). Distinct from existing datasets, most questions in SportsTime reflect real-world analytical demands, requiring models to integrate evidence from multiple temporally separated events rather than relying on a single moment. To support this, we further provide over 50,000 step-wise temporal evidence labels aligned with intermediate reasoning steps, enabling explicit supervision and fine-grained evaluation (see Fig. 2 for data examples). We employ a dual-track evaluation protocol: an LLM-as-Judge framework for open-ended QA, and a specialized step-wise grounding alignment (SGA) evaluation to assess reasoning quality, with details provided in Section 5.1.

To construct the benchmark at scale while maintaining domain fidelity, we adopt an expert-guided semi-automatic annotation pipeline. First, structured question templates for each sport are designed by domain experts, covering diverse tasks including perception, temporal, tactical, causal, and counterfactual reasoning. Then, these templates are instantiated by strong MLLMs to generate candidate QA samples and corresponding temporal evidence. Finally, peer reviewers and domain experts revise the annotations to ensure temporal accuracy, logical consistency, and domain correctness.

Building upon SportsTime, we further propose a **Chain-of-Time Reasoning (CoTR)** framework that formulates long-video understanding as a step-wise reasoning process over explicit temporal evidence. CoTR consists of two complementary mechanisms. First, during training, we introduce a temporal-reward Group Relative Policy Optimization (tr-GRPO) strategy that encourages models to align their reasoning steps with temporally grounded evidence, reinforcing correct localization and discouraging shortcuts. Second, during inference, we adopt a temporal search method based on an anchor-observe-infer loop. The model first anchors to candidate temporal segments, then iteratively observes retrieved clips to gather contextual evidence, and finally performs compositional reasoning to produce the final answer. This iterative evidence-seeking process enables explicit temporal grounding and improves long-horizon reasoning robustness.

Extensive experiments demonstrate that current MLLMs struggle with long-form compositional reasoning in the SportsTime benchmark, even strong proprietary models achieve limited accuracy. In contrast, the proposed CoTR method substantially improves both temporal localization and reasoning coherence by explicitly aligning intermediate reasoning steps with temporally grounded video evidence. These findings underscore the intrinsic difficulty of temporal compositional reasoning, demonstrating the necessity of benchmarks and methods that explicitly model dispersed temporal evidence to improve long-form video understanding. In summary, our contributions are threefold:

- We introduce SportsTime, a large-scale benchmark for long-form sports video understanding, featuring temporal compositional reasoning tasks and fine-grained step-wise temporal evidence annotations.
- We propose Chain-of-Time Reasoning (CoTR), an evidence-driven method that enforces temporal grounding during training through reward-aligned op-

timization and enables robust long-horizon reasoning via an evidence-seeking inference loop.
- Extensive experiments demonstrate that SportsTime reveals substantial limitations of existing MLLMs in long-horizon reasoning, and that CoTR consistently improves both answer accuracy and step-wise temporal grounding alignment across diverse settings.

2 Related Work

2.1 Multimodal Sports Benchmarks

Existing multimodal sports benchmarks largely fall into two categories. (1) Single-sport efforts, most notably in soccer, have progressed from task-specific benchmarks to more unified soccer-specific modeling. SoccerNet [8] established representative tasks such as replay grounding and camera-shot segmentation, while later works including SoccerReplay-1988 [26] and SoccerMaster [45] scale up multimodal soccer data and support more unified soccer understanding pipelines. (2) Multi-sport benchmarks mainly broaden coverage across sports while moving from generic VideoQA toward richer perception and reasoning evaluation. Sports-QA [41] and SPORTU [42] evaluate multi-sport QA, rule understanding, and slow-motion video reasoning, whereas SportR [40] introduces CoT and spatial-grounding supervision. FineSports [43] provides annotations for action understanding and spatio-temporal localization. However, most benchmarks still emphasize short clips or fine-grained action labels, leaving long-form temporal compositional reasoning with step-wise verifiable evidence largely underexplored.

2.2 Temporally Grounded Video Understanding

Recent work has increasingly focused on temporal understanding in long videos. One line of work [9, 14, 22, 35, 39] focuses on temporal grounding, aligning a language query with the relevant temporal segment in a video, i.e., query-to-time alignment, often by predicting timestamps, temporal spans, or moment proposals [4, 13, 31–33, 47]. Another line of work [20, 29, 49, 55] investigates how temporally localized evidence can be used to support video understanding and reasoning. TOGA [14] jointly predicts an open-ended answer and the temporal spans that support the final response. In addition, LongVT [46] and VITAL [53] use temporal localization as part of a tool-augmented reasoning process over long videos. Unlike prior work, we formulate long-video understanding as chain-of-time reasoning, with step-wise temporal anchors serving as intermediate states that guide evidence acquisition and multi-step inference.

3 SportsTime Benchmark

This section introduces **SportsTime**, a long-form sports video understanding benchmark with fine-grained chain-of-time annotations, from its construction (Sec. 3.1) and characteristics (Sec. 3.2). Release details and ethical considerations are provided in Appendix A.1.

3.1 Benchmark Construction

Benchmark Formulation and Design. We formulate SportsTime as a long-form sports video reasoning benchmark with QA pairs and chain-of-time annotations. Each sample consists of a video, a question, an open-ended answer, and a step-wise reasoning chain in which each intermediate step is grounded in explicit temporal evidence. We highlight three key designs of SportsTime below. (1) **Open-ended question answering.** We adopt open-ended QA rather than multiple-choice QA because it better reflects real-world usage and is also more challenging for models. (2) **Five task types for broad coverage.** We organize the questions into five task types, including causal, tactical, counterfactual, temporal, and perception reasoning. These categories cover a broad spectrum of abilities, from visual recognition to higher-level game understanding. (3) **Fine-grained Chain-of-Time annotation.** A key feature of SportsTime is that we annotate each reasoning step with a supporting timestamp or time span, making the reasoning path explicitly verifiable and enabling direct supervision for temporally compositional long video reasoning.

Data Collection. To ensure diversity across sports, we collect official broadcast videos from five team sports, including American Football, Ice Hockey, Soccer, Basketball, and Volleyball, which differ substantially in camera conventions, editing styles, and event structures. In addition, we further include men's and women's games and international, professional, and collegiate events to introduce structured diversity in gameplay and presentation. SportsTime includes 1,575 videos, with 208 full-match videos and 1,367 highlight videos. We select full matches to benchmark long-horizon reasoning under sparse evidence, and highlights to improve coverage of event-dense and stylistically diverse scenarios.

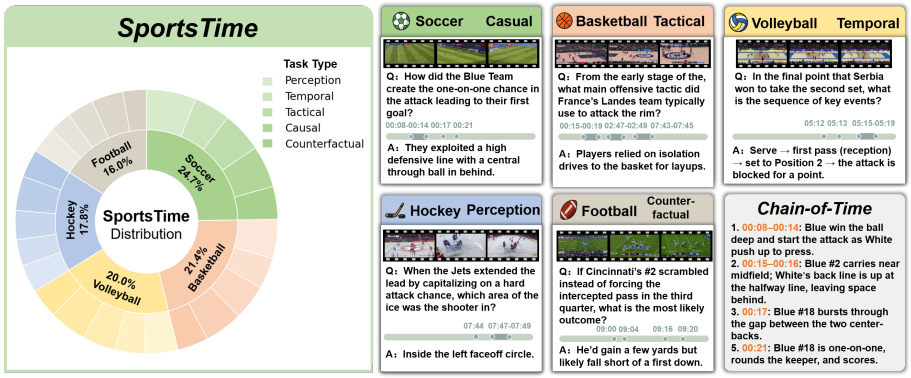


Fig. 2: Overview of the SportsTime benchmark covering five sports and five reasoning types, with Chain-of-Time examples.

Annotation Pipeline. Since long-form sports video QA samples simultaneously require large-scale coverage, temporally valid evidence, and domain-correct sports knowledge, neither purely manual construction nor purely automatic generation is sufficient. To address this challenge, we propose **an expert-guided semi-automatic annotation framework**. After data collection, the framework consists of three stages: expert template design, candidate QA generation, and two-stage manual review.

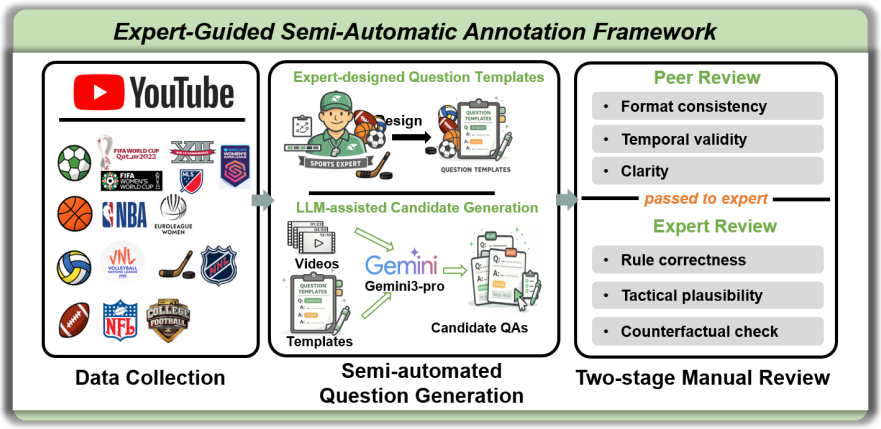


Fig. 3: Overview of our expert-guided semi-automatic annotation pipeline. We collect long-form sports videos, generate candidate QAs with expert-designed templates and LLM assistance, and ensure quality through two-stage manual review.

In the first stage, we invite sports experts to design question templates, which explicitly constrain the question types and semantic scope. In the second stage, we use an LLM to generate candidate QA samples based on the videos and templates, which improves the efficiency of large-scale data construction. This design turns open-ended generation into controlled generation, improving the usability and coverage of the candidate samples. Furthermore, to ensure data reliability, we introduce a two-stage manual review mechanism. Firstly, peer annotators conduct an initial check, focusing on format consistency, temporal validity, and clarity of expression, in order to filter out obviously low-quality samples. Secondly, domain experts perform final review, focusing on rule correctness, tactical plausibility, and the feasibility of counterfactual scenarios. This pipeline preserves the scalability of sample generation while effectively reducing temporal errors, semantic ambiguity, and domain-knowledge bias.

3.2 Benchmark Characteristics

Table 1 compares SportsTime with sports QA and video understanding benchmarks, and Fig. 4 summarizes the benchmark statistics of SportsTime. Com-

Table 1: Comparison with representative sports QA and video reasoning benchmarks. A and **M** indicate automatic and manual annotation types, respectively. **Stepwise Time** indicates temporal grounding aligned to reasoning steps: **S**= step-wise time spans, **TS**+**S**= step-wise timestamp and span.

Dataset	Type	Dur. (s)	#QA	Anno. Type	CoT	Time Anno.	Stepwise Time	QA Type
SportQA [41]	Sports	clips	70,000	A+M	✗	✗	✗	MCQ
Sports-QA [19]	Sports	15.0	94,000	A+M	✗	✗	✗	Open
SPORTU [42]	Sports	clips	12,048	A+M	✓	✗	✗	MCQ+Open
SPORTR [40]	Sports	4.96	20,000	M	✓	✗	✗	MCQ+Open
DeepSport [57]	Sports	–	84,700	A	✓	✗	✗	Open
LongVideoBench [38]	General	473.0	6,678	M	✗	✗	✗	MCQ
LVBench [34]	General	4101.0	1,549	M	✗	✓	✗	MCQ
Video-MME [11]	General	1017.9	2,700	M	✗	✗	✗	MCQ
MLVU [56]	Narrative	930	3,102	M	✗	✗	✗	MCQ
MINERVA [25]	General	743.23	1,515	M	✓	✓	✗	MCQ
VRBench [50]	Narrative	5,796.0	8,243	M	✓	✓	**S**	MCQ+Open
Ours	Sports	1053.25	14,326	M	✓	✓	**TS**+**S**	Open

pared with existing sports QA datasets (e.g., SportQA [41]), SportsTime targets longer videos and adopts an open-ended QA setting. This makes it better suited for evaluating compositional reasoning in realistic analysis scenarios. Compared with prior benchmarks that include CoT annotations (e.g., SPORTU [42], SPORTR [40], and DeepSport [57]), SportsTime further provides step-wise temporal annotations. This design makes the reasoning process verifiable and enables direct supervision and evaluation of intermediate reasoning paths. Compared with representative general long-video reasoning benchmarks (e.g., Video-MME, LongVideoBench, and MLVU, along with related long-video benchmarks [11, 25, 34, 38, 50, 56]), SportsTime focuses on sports scenarios and more systematically captures domain-specific challenges, including long-horizon event evolution, sparse key events, and cross-segment evidence dependencies. In particular, compared with VRBench [50], our sports setting requires much finer temporal grounding at each reasoning step. Each step is typically associated with an event timestamp or a short, second-scale span, rather than the longer, minute-scale spans that often suffice in narrative videos. This reflects the sparsity and momentary nature of decisive sports events, which makes precise evidence composition substantially harder. Overall, SportsTime provides a realistic and challenging testbed for broader research on long-video understanding.

4 Method

In this section, we present CoTR, our **C**hain-**o**f-**T**ime **R**easoning approach for temporal compositional reasoning. The core idea is to make the model reason with explicit temporal evidence. To reliably induce this behavior, we adopt a three-stage pipeline. First, we introduce a **visual time anchoring** step (Sec. 4.2) that turns temporal grounding into a directly observable visual cue. Second, we perform **reinforcement learning with temporally grounded reasoning** (Sec. 4.3) to encourage reasoning based on temporal evidence. Third, we intro-

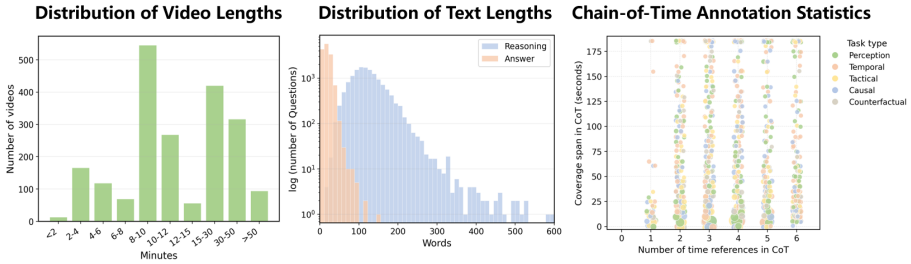


Fig. 4: Statistics of SportsTime. From left to right: video-length distribution, word-length distributions of reasoning chains and answers, and Chain-of-Time statistics.

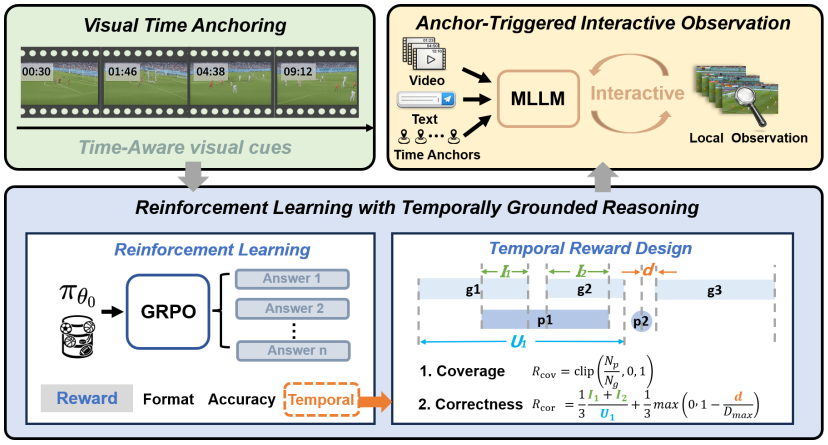


Fig. 5: Overview of the Chain-of-Time Reasoning (CoTR) Framework. CoTR consists of three stages: visual time anchoring, reinforcement learning with temporally grounded reasoning, and anchor-triggered interactive observation.

duce **anchor-triggered interactive observation** (Sec. 4.4) to iteratively verify and revise reasoning via anchor-based local clip retrieval. The overall framework of CoTR is illustrated in Fig. 5.

4.1 Problem Formulation

CoTR frames long-form video QA as a chain-of-time reasoning problem. Given a video V and a question q, the model predicts the answer a through a sequence of reasoning steps, each grounded in localized temporal evidence. Concretely, we define each step as a tuple $n_t = \langle s_t, \tau_t \rangle$, where s_t is a textual statement and τ_t is its supporting time anchor, represented as either a timestamp or a temporal span. The full trajectory becomes $\mathcal{T} = [n_1, \ldots, n_T, a]$, yielding the factorization

$$\pi_\theta(\mathcal{T} \mid V, q) = \Big(\prod_{t=1}^{T} \pi_\theta(n_t \mid V, q, n_{<t}) \Big) \cdot \pi_\theta(a \mid V, q, n_{\leq T}), \tag{1}$$

Under this formulation, each intermediate claim is paired with a retrievable temporal anchor, which enables iterative local evidence verification and revision.

4.2 Visual Time Anchoring

A practical challenge is that many MLLMs struggle to align events with their exact times in long videos. Because temporal progress is not directly observable, temporal anchors can drift even when the event itself is correctly identified. Inspired by the philosophy behind DeepSeekOCR [37] and Number it [39] that *a picture is worth a thousand words*, we introduce a simple but effective method. We burn an explicit timestamp overlay into the top-right corner of video frames with a consistent `mm:ss` format. This converts temporal grounding into a directly readable visual signal, allowing the model to infer time through visual-text recognition and reducing anchor drift in long contexts. Empirically, this simple method improves the learnability of time anchoring.

4.3 Reinforcement Learning with Temporally Grounded Reasoning

To improve temporal grounding, we optimize the policy with reinforcement learning to promote evidence-seeking behavior and encourage the model to ground its reasoning on time anchors. Starting from the initial backbone policy π_0, we optimize a trajectory policy $\pi_\theta(\mathcal{T} \mid V, q)$ over anchored trajectories $\mathcal{T} = [n_1, \ldots, n_T, a]$ using temporal-reward GRPO objective, which improves stability by comparing rollouts within a group. At each update, we sample multiple trajectories per (V, q), compute rewards, form group-relative advantages, and update π_θ to maximize the reward while preserving the anchored generation protocol.

Reward design. Our total reward is a weighted sum of three components,

$$R(\mathcal{T}) = \lambda_{\text{fmt}}\, r_{\text{fmt}}(\mathcal{T}) + \lambda_{\text{acc}}\, r_{\text{acc}}(\mathcal{T}) + \lambda_{\text{temporal}}\, r_{\text{temporal}}(\mathcal{T}), \tag{2}$$

where r_{fmt} is a binary structural reward that checks whether the output follows the required format, and r_{acc} is a task-level answer correctness reward. Our main focus here is r_{temporal}, which evaluates the quality of temporal grounding. Concretely, our temporal reward consists of two parts—*coverage* and *correctness*:

$$r_{\text{temporal}}(\mathcal{T}) = \alpha\, r_{\text{cov}}(\mathcal{T}) + (1-\alpha)\, r_{\text{cor}}(\mathcal{T}), \tag{3}$$

where α is a tunable weight.

(1) Step-wise coverage. r_{cov} measures step-level temporal anchor coverage by rewarding trajectories in which reasoning steps are explicitly grounded in time. Concretely, we segment the model's `<thinking>` into steps and compute the proportion of steps that contain at least one explicit time anchor.

(2) Temporal correctness. r_{cor} measures how well predicted anchors align with ground-truth anchors. We extract predicted anchors from the model's `<thinking>` and obtain ground-truth step anchors from the supervised process annotations. For each ground-truth anchor, we compute its best-match score against the predicted anchors, using span IoU for span–span matches and a distance-aware similarity for point–span or point–point matches. We then average these scores over all ground-truth anchors to obtain $r_{\text{cor}} \in [0, 1]$. Detailed matching rules and scoring definitions are provided in Appendix B.2.

This reward design provides explicit supervision for temporal grounding: r_{cov} encourages the model to cover all required evidence, while r_{cor} pushes predicted anchors toward the annotated evidence spans. As a result, tr-GRPO makes time anchors more learnable intermediate states, encouraging the policy to generate reasoning chains whose anchors are better aligned with verifiable evidence.

4.4 Anchor-Triggered Interactive Observation

Once the model has learned to produce time-anchored reasoning steps, we further use these anchors as actionable cues for test-time observation. Given a video V and a question q, the model generates an anchored reasoning trajectory $\mathcal{T} = [n_1, \ldots, n_T, a]$ with $n_t = \langle s_t, \tau_t \rangle$, where each step s_t is accompanied by a temporal anchor τ_t. We treat each predicted τ_t as an explicit temporal query for local evidence retrieval in the corresponding neighborhood of V.

Anchor-triggered local sampling. For a point anchor $\tau_t = \texttt{mm:ss}$, we sample a short temporal window centered at τ_t; for a span anchor $\tau_t = [t_t^s, t_t^e]$, we sample multiple clips uniformly from within the span. Each anchor is thus converted into a set of local clips $\{c_t^{(j)}\}_{j=1}^{J_t}$, where each clip consists of L frames sampled at a fixed stride. This anchor-triggered sampling replaces global scanning of a long video with bounded local observation around predicted evidence locations.

Evidence-grounded reasoning. We then perform reasoning over the anchored steps. At turn t, the model is given the question q, the current step s_t, the anchor τ_t, and the retrieved local clips $\{c_t^{(j)}\}_{j=1}^{J_t}$ as visual evidence. Based on this local evidence, the model verifies and revises the current step before proceeding to the next turn. Repeating this process yields a refined trajectory $\widetilde{\mathcal{T}}$ whose intermediate claims are explicitly checked against retrieved video content. Finally, the model outputs the final answer a conditioned on the accumulated verified evidence across turns.

5 Experiments

This section first introduces the experimental setting in Sec. 5.1, and then presents the main results in Sec. 5.2, followed by ablation studies in Sec. 5.3.

Table 2: Performance on SportsTime. "Visual Input" denotes the video input budget used for each model, reported either as the maximum number of sampled frames or the sampling rate. All scores are in %.

Model	Visual Input	SportsTime					
		Perception	Temporal	Tactical	Causal	Counterfactual	Avg.
Proprietary							
GPT-5	0.2 fps	38.77	27.45	43.85	44.77	46.93	40.72
Gemini-2.5-Pro	0.2 fps	39.66	19.78	29.49	39.41	13.71	29.37
Open-source							
Qwen3-VL-8B-Instruct [2]	768 frames	24.27	14.31	33.26	23.08	44.47	27.45
Qwen3-VL-4B-Instruct	768 frames	23.26	13.14	31.11	22.92	34.17	25.15
VideoLLaMA3-7B [52]	768 frames	20.06	10.98	16.92	14.00	28.54	17.96
InternVideo2.5-8B [5]	512 frames	18.12	10.39	19.83	9.85	26.21	16.68
Video-R1-7B [10]	768 frames	19.26	8.04	23.25	18.62	33.01	20.40
MiniCPM-V4.5-8B [48]	512 frames	16.38	10.14	13.50	16.18	27.27	16.48
GLM-4.6v-Flash-9B [15]	640 frames	2.27	1.18	3.76	3.54	13.20	4.62

5.1 Experimental Settings

Implementation Details. We sample 128 frames per video for training and up to 768 frames for inference. We use Qwen3-VL-4B as the backbone model. All experiments are conducted on 2× NVIDIA H100 (80GB). Additional implementation details, full GRPO hyperparameters, reward definitions, and the anchor-extraction parser used for Fig. 6(a) are provided in Appendix A.3.

Dual-Track Evaluation Protocol. We consider two settings for evaluation.

Open-ended QA. We follow an LLM-as-Judge protocol and report all main results using a fixed judge (Qwen2.5-VL-7B [3]). To assess the reliability of this, we further conduct a reliability study (Sec. 5.2) by comparing judgments from two additional LLM judges (MiniMax-M2.5 and GLM-4.7 [51]) and human raters.

Step-wise Grounding Alignment (SGA) Evaluation We employ both objective and subjective assessments. Objectively, we measure temporal alignment to annotated evidence using temporal IoU between predicted time and ground-truth. Subjectively, we conduct human evaluation to assess (i) whether the evidence genuinely supports each reasoning step, (ii) whether intermediate conclusions are reasonable, and (iii) whether the overall chain is faithful and logically consistent.

5.2 Main Results

Results on SportsTime. In Table 2, we report performance on SportsTime across five task types and the overall average. Overall, the absolute accuracy is still far from solved: even strong proprietary models reach only 40.72% (GPT-5) and 29.37% (Gemini-2.5-Pro), while open-source baselines are substantially lower. Such a low performance ceiling is expected because SportsTime demands

Table 3: Effectiveness of CoTR on SportsTime. All scores are in %. Best and second-best results are highlighted by **bold** and gray underline, respectively. The last column reports **human evaluation** on a stratified subset of 200 test examples. **Improvement** denotes the absolute gain over *the baseline model* (Qwen3-VL-4B-Instruct), measured in percentage points.

Model	SportsTime						
	Perception	Temporal	Tactical	Causal	Counterfactual	Avg.	Human Avg.(subset)
Qwen3-VL-4B-Instruct	23.26	13.14	31.11	22.92	34.17	25.15	24.50
Qwen2.5-VL-3B-Instruct	16.67	7.06	27.18	12.31	22.52	17.16	17.50
InternVL2.5-4B [5]	21.99	14.83	30.58	16.30	39.23	24.57	26.00
MiniCPM-v4-4B [48]	14.21	10.36	19.20	12.08	26.06	16.23	16.50
Ovis2-4B [24]	11.87	6.34	13.96	6.76	5.25	9.02	8.00
LLaVA-OneVision1.5-4B [1]	13.88	7.40	13.61	7.73	16.57	11.81	12.50
Ours							
Ours(tr-GRPO)	22.01	13.73	**35.73**	25.54	39.81	27.31	29.00
Ours(full)	**26.42**	**15.43**	34.55	**27.21**	42.22	**29.24**	**30.50**
Improvement	(+3.16% ↑)	(+2.29% ↑)	(+3.44% ↑)	(+4.29% ↑)	(+8.05% ↑)	(+4.09% ↑)	(+6.00% ↑)

temporal compositional reasoning over long-form matches, where decisive evidence is sparse, brief, and widely separated in time. Under a fixed frame budget, models may fail to retrieve the key moments and instead fall back to semantic priors, which particularly hurts Temporal and Causal questions that demand evidence-based linking across distant events.

Effectiveness of CoTR. In Table 3, building on Qwen3-VL-4B, our CoTR yields consistent gains. **CoTR (full)** improves the overall average from 25.15% to 29.24%, with notable improvements on Counterfactual and Causal. After tr-GRPO fine-tuning, the model shows a pronounced gain on Tactical, which may primarily reflect improved command of sports-specific tactical concepts, and is potentially further aided by the temporal reward. With full CoTR, we improve both Perception (+3.16%) and Causal (+4.29%), suggesting that our method improves both the identification of relevant visual evidence and the integration of temporally dispersed clues for causal attribution. To validate the reliability of LLM-as-Judge, we additionally perform human evaluation on 200 stratified test samples, and find that its ranking is highly consistent with human judgments.

SGA Evaluation. We evaluate the quality of reasoning chains in two ways.

(1) Objective Evaluation. Fig. 6(a) reports both final-answer accuracy and temporal evidence quality. Zero-shot CoT achieves only modest accuracy and, more importantly, rarely grounds its reasoning with explicit temporal spans (Anchor 19.49%). In contrast, time-prompted CoT drastically increases the frequency of span outputs, but the spans are poorly aligned with the true evidence (mIoU 0.12, Hit@0.5 12.12%). This gap indicates that simply "asking for timestamps" encourages format compliance rather than evidence faithfulness, i.e., the model may attach arbitrary or weakly related time windows to justify intermediate claims, resulting in temporal drift with superficially grounded rationales.

(a) Accuracy and temporal evidence quality.

Model	Acc(%)↑	Anchor(%)↑	mIoU↑	H@0.5(%)↑
Base	25.15	11.95	0.2601	30.03
Zero-shot CoT [16]	24.01	19.49	0.2042	26.34
Time-prompted CoT	19.53	91.80	0.1244	12.12
Native-GRPO [30]	26.72	13.10	0.3128	32.09
tr-GRPO(Ours)	27.31	95.76	0.5812	56.89

(b) Human assessment under SGA.

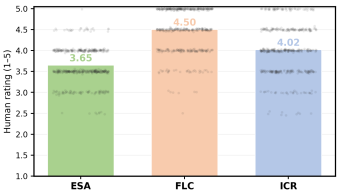


Fig. 6: SGA Evaluation. "mIoU" is the mean span IoU between predicted and reference time windows. "H@τ" is the fraction of examples whose span IoU exceeds threshold τ. Dots correspond to ratings of individual samples, while bars indicate the mean score.

Our method closes this gap by explicitly internalizing the desired behavior into the reward design. In particular, tr-GRPO optimizes chain-of-time generation with rewards that jointly encourage temporal evidence coverage and correctness. Compared with native-GRPO, which improves answer accuracy but still exhibits weak temporal grounding, our reward design yields a fundamentally better trade-off between reasoning outcome and evidence quality. Ours reaches 27.31% accuracy while maintaining high anchoring coverage (Anchor 95.76%) and substantially stronger temporal alignment (mIoU 0.58, Hit@0.5 56.89%). The large gains indicate that the improvement is not merely better answer matching, but a genuine reduction in temporal drift, as the predicted evidence more reliably support reasoning steps with the correct moments.

(2) Human Assessment under SGA. Fig. 6(b) reports human evaluation results. We randomly sample 100 examples and ask five raters to score the chains generated by CoTR on three dimensions: evidence-span accuracy (ESA), intermediate claim reliability (ICR), and faithful and logical consistency (FLC). CoTR achieves a high score on FLC (4.50/5), which indicates coherent and logically consistent chains, and a high ICR (4.02/5), suggesting largely reliable intermediate claims. ESA reaches 3.65/5, showing that although temporal spans generally support the corresponding steps, fine-grained localization remains improvable.

Comparison on Other Benchmarks. To evaluate generalization beyond SportsTime, we further test CoTR on several public long video reasoning benchmarks [11,34,56] in Table 4. CoTR achieves consistently competitive performance across these datasets, showing promising transferability beyond SportsTime.

Reliability of LLM-as-Judge. We use three independent LLM judges: GLM-4.7, MiniMax-M2.5, and Qwen2.5-VL-7B, together with human raters for open-ended evaluation on SportsTime. As shown in Table 6, the judges achieve high average pairwise agreement, while Fleiss' κ indicates moderate inter-judge consistency. In addition, individual LLM judges show moderate agreement with human judgments. These results support the use of LLM-as-Judge for scalable evaluation, although there remains room for improvement.

Table 4: Comparison on general video benchmarks. All scores are in %. **Improvement** denotes the absolute gain over *the baseline model* (Qwen3-VL-4B-Instruct).

Models	LVBench	MLVU	VideoMME
Qwen3-VL-8B-Instruct	58.0	**78.1**	71.9
VideoLLaMA3-7B	44.3	68.7	61.1
InternVideo2.5-8B	46.4	72.8	65.1
Video-R1-7B	35.4	58.9	59.3
InternVL2.5-4B	40.8	48.3	63.6
SF-LLaVA-1.5-3B [44]	43.3	68.8	49.2
Qwen2.5-VL-3B-Instruct	43.3	68.2	61.5
Qwen3-VL-4B-Instruct	56.2	75.3	69.3
Ours	**59.9**	76.2	**72.1**
Improvement	(+3.7% ↑)	(+0.9% ↑)	(+2.8% ↑)

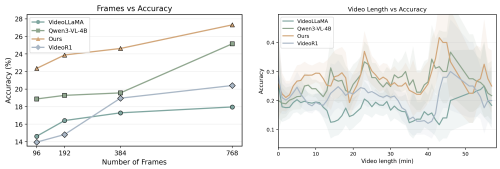
Table 5: Ablation of our method. All scores are in %. The last two rows correspond to the main components of our method.

Setting	SportsTime		
	Temporal	Tactical	All
Base	13.14	31.11	25.15
+ SFT	6.98	19.90	17.61
+ Native-GRPO	11.76	32.65	26.72
+ tr-GRPO w/o ts	13.14	32.21	26.12
+ tr-GRPO w/ ts	13.73	35.73	27.31
+ **AT-IO (Ours)**	**15.43**	**34.55**	**29.24**

Table 6: Open-ended QA accuracy under multiple judges. All scores are in %.

Model	Qwen	MiniMax	GLM [51]	Human
InternVideo2.5-8B	16.68	14.25	17.94	17.37
Qwen3-VL-4B	25.15	24.01	25.89	24.08
Ours-4B	29.24	28.74	30.23	29.60

Judge consistency:
Avg. pairwise agreement = 88.34%, Fleiss' $\kappa = 0.57$.
Human alignment:
Cohen's κ (Qwen/MiniMax/GLM vs Human) = 0.6467/0.5759/0.5882.



(a) Frames vs. Acc. **(b)** Video length vs. Acc.

Fig. 7: Video setting ablation studies. (a) Accuracy as a function of frame budget. (b) Accuracy as a function of video length.

5.3 Ablation Studies

Method Ablation. We conduct ablations to assess our method (Table 5). A direct SFT baseline performs worse than the base model, with frequent repetitive generation on many samples. We attribute this to the structural variability and non-uniqueness of temporal reasoning chains, which make direct token-level imitation brittle and potentially harder to optimize at the 4B scale. Native-GRPO brings only a small gain and even reduces Temporal, suggesting that generic RL fine-tuning is insufficient. In contrast, tr-GRPO yields larger improvements, with a notable boost on Tactical, highlighting the role of our temporal-reward design. Interestingly, removing the timestamp-overlay (w/o ts) causes a clear drop, indicating that visually accessible timestamps help the model better leverage temporal evidence. Finally, adding AT-IO (**A**nchor-**T**riggered **I**nteractive **O**bservation) achieves the best overall performance, showing that test-time verification provides complementary benefits beyond reward-based training.

Video Setting Ablation. Fig. 7(a) shows that accuracy generally improves with larger frame budgets, but the gain is model-dependent. Fig. 7(b) reveals a surprising pattern: accuracy does not decrease monotonically with video length, so length alone is not the decisive factor. Instead, difficulty is likely driven by

confounding factors such as video type and event structure. Notably, our method retains an advantage in longer videos, indicating its effectiveness.

6 Conclusion

In this work, we introduce **SportsTime**, a large-scale benchmark for long-form sports video understanding with step-wise temporal evidence annotations, and propose **Chain-of-Time Reasoning (CoTR)**, a framework that unifies temporally grounded training and inference for long video reasoning. Our results show that explicit temporal evidence supervision and step-wise evidence-seeking substantially improve both temporal compositional reasoning and grounding quality. We hope SportsTime will provide a strong testbed for future research on long-horizon, evidence-grounded video understanding.

Acknowledgements

This research is supported by Key Science & Technology Project of Anhui Province under Grant No. 202523j08050027 and the National Key Research and Development Program of China under Grant No. 2025YFE0216600.

References

1. An, X., Xie, Y., Yang, K., Zhang, W., Zhao, X., Cheng, Z., Wang, Y., Xu, S., Chen, C., Zhu, D., et al.: Llava-onevision-1.5: Fully open framework for democratized multimodal training. arXiv preprint arXiv:2509.23661 (2025)
2. Bai, S., Cai, Y., Chen, R., Chen, K., Chen, X., Cheng, Z., Deng, L., Ding, W., Gao, C., Ge, C., et al.: Qwen3-vl technical report. arXiv preprint arXiv:2511.21631 (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., et al.: Qwen2.5-vl technical report (2025)
4. Chen, G., Liu, Y., Huang, Y., He, Y., Pei, B., Xu, J., Wang, Y., Lu, T., Wang, L.: Cg-bench: Clue-grounded question answering benchmark for long video understanding. arXiv preprint arXiv:2412.12075 (2024)
5. Chen, Z., Wang, W., Cao, Y., Liu, Y., Gao, Z., Cui, E., Zhu, J., Ye, S., Tian, H., Liu, Z., et al.: Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling. arXiv preprint arXiv:2412.05271 (2024)
6. Clark, C., Zhang, J., Ma, Z., Park, J.S., Salehi, M., Tripathi, R., Lee, S., Ren, Z., Kim, C.D., Yang, Y., et al.: Molmo2: Open weights and data for vision-language models with video understanding and grounding. arXiv preprint arXiv:2601.10611 (2026)
7. Cui, Y., Zeng, C., Zhao, X., Yang, Y., Wu, G., Wang, L.: Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In: ICCV. pp. 9921–9931 (October 2023)

8. Deliege, A., Cioppa, A., Giancola, S., Seikavandi, M.J., Dueholm, J.V., Nasrollahi, K., Ghanem, B., Moeslund, T.B., Van Droogenbroeck, M.: Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In: CVPRW. pp. 4508–4519 (June 2021)
9. Ding, Y., Cao, S., Jiao, L., Li, Y., Wang, Z., Liu, Z., Zhang, L.: Retrieving any relevant moments: Benchmark and models for generalized moment retrieval (2026), https://arxiv.org/abs/2605.02623
10. Feng, K., Gong, K., Li, B., Guo, Z., Wang, Y., Peng, T., Wu, J., Zhang, X., Wang, B., Yue, X.: Video-r1: Reinforcing video reasoning in mllms. arXiv preprint arXiv:2503.21776 (2025)
11. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., Chen, P., Li, Y., Lin, S., Zhao, S., Li, K., Xu, T., Zheng, X., Chen, E., Shan, C., He, R., Sun, X.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: CVPR. pp. 24108–24118 (June 2025)
12. Ghasemzadeh, S.A., Van Zandycke, G., Istasse, M., Sayez, N., Moshtaghpour, A., De Vleeschouwer, C.: Deepsportlab: a unified framework for ball detection, player instance segmentation and pose estimation in team sports scenes. arXiv preprint arXiv:2112.00627 (2021)
13. Guo, Y., Liu, J., Li, M., Cheng, D., Tang, X., Sui, D., Liu, Q., Chen, X., Zhao, K.: Vtg-llm: Integrating timestamp knowledge into video llms for enhanced video temporal grounding. AAAI **39**(3), 3302–3310 (Apr 2025)
14. Gupta, A., Roy, A., Chellappa, R., Bastian, N.D., Velasquez, A., Jha, S.: Toga: Temporally grounded open-ended video qa with weak supervision. In: ICCV. pp. 23593–23603 (October 2025)
15. Hong, W., Yu, W., Gu, X., Wang, G., Gan, G., Tang, H., Cheng, J., Qi, J., Ji, J., Pan, L., et al.: Glm-4.5 v and glm-4.1 v-thinking: Towards versatile multimodal reasoning with scalable reinforcement learning. arXiv preprint arXiv:2507.01006 (2025)
16. Kojima, T., Gu, S.S., Reid, M., Matsuo, Y., Iwasawa, Y.: Large language models are zero-shot reasoners. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) NeurIPS. vol. 35, pp. 22199–22213. Curran Associates, Inc. (2022)
17. Lee, K., Kim, E., Choi, J., Chang, B.: Noah: Benchmarking narrative prior driven hallucination and omission in video large language models. arXiv preprint arXiv:2511.06475 (2025)
18. Li, C., Im, E.W., Fazli, P.: Vidhalluc: Evaluating temporal hallucinations in multimodal large language models for video understanding. In: CVPR. pp. 13723–13733 (June 2025)
19. Li, H., Deng, A., Ke, Q., Liu, J., Rahmani, H., Guo, Y., Schiele, B., Chen, C.: Sports-qa: A large-scale video question answering benchmark for complex and professional sports. arXiv preprint arXiv:2401.01505 (2024)
20. Li, R., Wang, X., Zhang, Y., Zohar, O., Wang, Z., Yeung-Levy, S.: Temporal preference optimization for long-form video understanding. arXiv preprint arXiv:2501.13919 (2025)
21. Li, Y., Xiao, J., Feng, C., Wang, X., Chua, T.S.: Discovering spatio-temporal rationales for video question answering. In: ICCV. pp. 13869–13878 (October 2023)
22. Liu, H., Ma, X., Zhong, C., Zhang, Y., Lin, W.: Timecraft: Navigate weakly-supervised temporal grounded video question answering via bi-directional reasoning. In: ECCV. pp. 92–107. Springer (2024)

23. Lu, H., Wang, J., Zhang, Y., Wang, R., Zheng, X., Tang, Y., Lin, D., Lu, L.: Elv-halluc: Benchmarking semantic aggregation hallucinations in long video understanding. arXiv preprint arXiv:2508.21496 (2025)
24. Lu, S., Li, Y., Xia, Y., Hu, Y., Zhao, S., Ma, Y., Wei, Z., Li, Y., Duan, L., Zhao, J., et al.: Ovis2. 5 technical report. arXiv preprint arXiv:2508.11737 (2025)
25. Nagrani, A., Menon, S., Iscen, A., Buch, S., Mehran, R., Jha, N., Hauth, A., Zhu, Y., Vondrick, C., Sirotenko, M., Schmid, C., Weyand, T.: Minerva: Evaluating complex video reasoning. In: ICCV. pp. 23968–23978 (October 2025)
26. Rao, J., Wu, H., Jiang, H., Zhang, Y., Wang, Y., Xie, W.: Towards universal soccer video understanding. In: CVPR. pp. 8384–8394 (June 2025)
27. Rao, J., Wu, H., Liu, C., Wang, Y., Xie, W.: MatchTime: Towards automatic soccer game commentary generation. In: Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing. pp. 1671–1685 (Nov 2024)
28. Rawal, R., Shirkavand, R., Huang, H., Somepalli, G., Goldstein, T.: Argus: Hallucination and omission evaluation in video-llms. In: ICCV. pp. 20280–20290 (October 2025)
29. Ren, S., Chen, S., Li, S., Sun, X., Hou, L.: TESTA: Temporal-spatial token aggregation for long-form video-language understanding. In: Findings of the Association for Computational Linguistics: EMNLP 2023. pp. 932–947. Association for Computational Linguistics, Singapore (Dec 2023)
30. Shao, Z., Wang, P., Zhu, Q., Xu, R., Song, J., Bi, X., Zhang, H., Zhang, M., Li, Y., Wu, Y., et al.: Deepseekmath: Pushing the limits of mathematical reasoning in open language models. arXiv preprint arXiv:2402.03300 (2024)
31. Sugandhika, C., Li, C., Rajan, D., Fernando, B.: Know-show: Benchmarking video-language models on spatio-temporal grounded reasoning. arXiv preprint arXiv:2512.05513 (2025)
32. Wang, H., Xu, Z., Cheng, Y., Diao, S., Zhou, Y., Cao, Y., Wang, Q., Ge, W., Huang, L.: Grounded-videollm: Sharpening fine-grained temporal grounding in video large language models. arXiv preprint arXiv:2410.03290 (2024)
33. Wang, S., Chen, G., Huang, D.a., Li, Z., Li, M., Li, G., Alvarez, J.M., Zhang, L., Yu, Z.: Videoitg: Multimodal video understanding with instructed temporal grounding. arXiv preprint arXiv:2507.13353 (2025)
34. Wang, W., He, Z., Hong, W., Cheng, Y., Zhang, X., et al.: Lvbench: An extreme long video understanding benchmark. In: ICCV. pp. 22958–22967 (October 2025)
35. Wang, Y., Wang, Z., Xu, B., Du, Y., Lin, K., Xiao, Z., Yue, Z., Ju, J., Zhang, L., Yang, D., et al.: Time-r1: Post-training large vision language model for temporal video grounding. arXiv preprint arXiv:2503.13377 (2025)
36. Wang, Y., Wang, Y., Zhao, D., Xie, C., Zheng, Z.: Videohallucer: Evaluating intrinsic and extrinsic hallucinations in large video-language models. arXiv preprint arXiv:2406.16338 (2024)
37. Wei, H., Sun, Y., Li, Y.: Deepseek-ocr: Contexts optical compression. arXiv preprint arXiv:2510.18234 (2025)
38. Wu, H., Li, D., Chen, B., Li, J.: Longvideobench: A benchmark for long-context interleaved video-language understanding. In: NeurIPS. vol. 37, pp. 28828–28857 (2024)
39. Wu, Y., Hu, X., Sun, Y., Zhou, Y., Zhu, W., et al.: Number it: Temporal grounding videos like flipping manga. In: CVPR. pp. 13754–13765 (June 2025)
40. Xia, H., Ge, H., Zou, J., Choi, H.W., Zhang, X., Suradja, D., Rui, B., Tran, E., Jin, W., Ye, Z., et al.: Sportr: A benchmark for multimodal large language model reasoning in sports. arXiv preprint arXiv:2511.06499 (2025)

41. Xia, H., Yang, Z., Wang, Y., Tracy, R., Zhao, Y., Huang, D., Chen, Z., Zhu, Y., Wang, Y.f., Shen, W.: Sportqa: A benchmark for sports understanding in large language models. In: Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers). pp. 5061–5081 (2024)
42. Xia, H., Yang, Z., Zou, J., Tracy, R., Wang, Y., Lu, C., Lai, C., He, Y., Shao, X., Xie, Z., et al.: Sportu: A comprehensive sports understanding benchmark for multimodal large language models. arXiv preprint arXiv:2410.08474 (2024)
43. Xu, J., Zhao, G., Yin, S., Zhou, W., Peng, Y.: FineSports: A multi-person hierarchical sports video dataset for fine-grained action understanding. In: CVPR. pp. 21773–21782. IEEE, Seattle, WA, USA (Jun 2024)
44. Xu, M., Gao, M., Li, S., Lu, J., Gan, Z., Lai, Z., Cao, M., Kang, K., Yang, Y., Dehghan, A.: Slowfast-llava-1.5: A family of token-efficient video large language models for long-form video understanding. arXiv preprint arXiv:2503.18943 (2025)
45. Yang, H., Rao, J., Wu, H., Xie, W.: Soccermaster: A vision foundation model for soccer understanding. arXiv preprint arXiv:2512.11016 (2025)
46. Yang, Z., Wang, S., Zhang, K., Wu, K., Leng, S., Zhang, Y., Li, B., Qin, C., Lu, S., Li, X., et al.: Longvt: Incentivizing" thinking with long videos" via native tool calling. arXiv preprint arXiv:2511.20785 (2025)
47. Yang, Z., Yu, Y., Zhao, Y., Lu, S., Bai, S.: Timeexpert: An expert-guided video llm for video temporal grounding. In: ICCV. pp. 24286–24296 (October 2025)
48. Yao, Y., Yu, T., Zhang, A., Wang, C., Cui, J., et al.: Minicpm-v: A gpt-4v level mllm on your phone. arXiv preprint arXiv:2408.01800 (2024)
49. Ye, J., Wang, Z., Sun, H., Chandrasegaran, K., Durante, Z., et al.: Re-thinking temporal search for long-form video understanding. In: CVPR. pp. 8579–8591 (June 2025)
50. Yu, J., Wu, Y., Chu, M., Ren, Z., Huang, Z., Chu, P., Zhang, R., He, Y., Li, Q., Li, S., Li, Z., Tu, Z., He, C., Qiao, Y., Wang, Y., Wang, Y., Wang, L.: Vrbench: A benchmark for multi-step reasoning in long narrative videos. In: ICCV. pp. 21655–21666 (October 2025)
51. Zeng, A., Lv, X., Zheng, Q., Hou, Z., Chen, B., Xie, C., Wang, C., Yin, D., Zeng, H., Zhang, J., et al.: Glm-4.5: Agentic, reasoning, and coding (arc) foundation models. arXiv preprint arXiv:2508.06471 (2025)
52. Zhang, H., Li, X., Bing, L.: Video-LLaMA: An instruction-tuned audio-visual language model for video understanding. In: Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations. pp. 543–553 (Dec 2023)
53. Zhang, H., Gu, X., Li, J., Ma, C., Bai, S., Zhang, C., Zhang, B., Zhou, Z., He, D., Tang, Y.: Thinking with videos: Multimodal tool-augmented reinforcement learning for long video reasoning. arXiv preprint arXiv:2508.04416 (2025)
54. Zhang, J., Jiao, Y., Chen, S., Zhao, N., Tan, Z., et al.: Eventhallusion: Diagnosing event hallucinations in video llms. arXiv preprint arXiv:2409.16597 (2024)
55. Zhou, B., Andonian, A., Oliva, A., Torralba, A.: Temporal relational reasoning in videos. In: ECCV (September 2018)
56. Zhou, J., Shu, Y., Zhao, B., Wu, B., Liang, Z., et al.: Mlvu: Benchmarking multi-task long video understanding. In: CVPR. pp. 13691–13701 (June 2025)
57. Zou, J., Xia, H., Ye, Z., Zhang, S., Lai, C., Ordonez, V., Shen, W., Chen, H.: Deepsport: A multimodal large language model for comprehensive sports video reasoning via agentic reinforcement learning. arXiv preprint arXiv:2511.12908 (2025)