SCBench: A Sports Commentary Benchmark for Video LLMs

Anonymous ACL submission

Abstract

Recently, significant advances have been made in Video Large Language Models (Video LLMs) in both academia and industry. However, methods to evaluate and benchmark the performance of different Video LLMs, especially their long-horizon, fine-grained video understanding capabilities, remain very limited. On one hand, current long video benchmarks use semantically simple videos (e.g., subtitled movies) where models understand the plot by processing only a few frames. On the other hand, these datasets lack diversity in task format, comprising only QA or multiple-choice, overlooking the models' text generation capabilities. Sports videos, which feature intricate visual information, sequential events, and emotionally charged commentary, present a critical challenge for Video LLMs, making sports commentary an ideal benchmarking task. Inspired by these challenges, we propose a novel task: sports video commentary generation and develop **SCBench** for Video LLMs. To construct such a benchmark, we introduce (1) a six-dimensional metric specifically designed for our task, upon which we propose a GPT-based evaluation method, and (2) **CommentarySet**, a dataset consisting of 5,775 annotated video clips and ground-truth labels tailored to our metric. Based on SCBench, we conduct comprehensive evaluations on multiple Video LLMs (e.g. VILA, Video-LLaVA, etc.). Our results show that InternVL-Chat-2 achieves the best performance of 5.44, surpassing the second-best by 1.04. Our work provides a fresh perspective for future research, aiming to enhance models' overall capabilities in complex, long-horizon visual understanding tasks. Code and dataset are available here.

1 Introduction

With the advancement of Video Large Language Models (Video-LLMs), neural networks are capable of processing video sequences to perform multi-modality tasks based on temporal and spatial information. Recent advancements have facilitated significant improvements in the automatic generation of video content summaries, captions, and even interactive commentary (Miech et al., 2019; Zhou et al., 2024; Lin et al., 2024; Liu et al., 2024a).

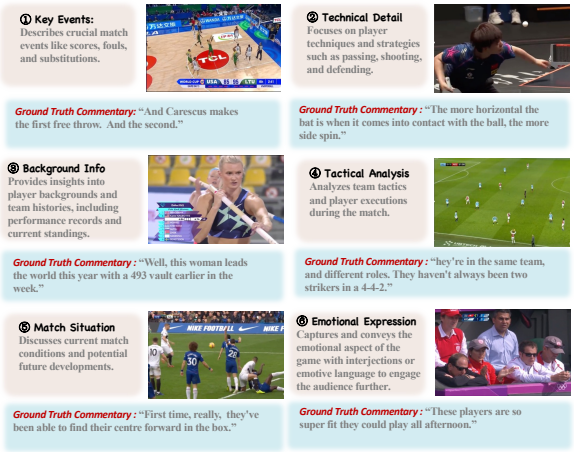

Figure 1: **A sample from Six-Dimensional Metric and CommentarySet**, including six dimensions' description, frames from videos, and corresponding ground truth commentary.

To holistically assess their margin of improvement, many benchmarks and evaluation methods have been proposed (Ning et al., 2023a; Sanders and Van Durme, 2024; Ning et al., 2023b; Fu et al., 2024). Existing benchmarks for Video-LLMs predominantly concentrate on general video understanding tasks(Ning et al., 2023a; Sanders and Van Durme, 2024), such as action recognition (Mangalam et al., 2023) and long context video understanding (Wang et al., 2024; Fang et al., 2024). However, these methods mostly capture image-level understanding tasks that mostly test the model's abilities in the spatial understanding tasks in the form of QA. Consequently, (Li et al., 2024) have curated tasks that require reasoning spanning over the entire videos for MLLMs. However, this benchmark still suffers from the lack

of dense events, where a 1 FPS frame-sampling rate is sufficient to perform well, meaning that the model's dense temporal understanding ability remains largely untested. Additionally, there is a lack of diversity in task format, comprising only QA or multi-choice QA, which overlooks the models' capacity for generating in-depth and precise texts, thus precluding a deeper evaluation of the models' analytical and interpretative abilities.

Discovering limitations in existing datasets, we turn to professional sports videos to present a challenging domain for video understanding for Video-LLMs. These videos involve complex human-object interactions, strategic gameplay, and visually demanding tasks such as event and motion spotting. The rapid shifts between high-action sequences and static moments require models to prioritize information effectively, while the need to interpret intricate motions and maintain context over time adds to the complexity, especially for real-time sports commentary.

Although many existing sports video datasets have been explored, they are not well-suited for evaluating video-based multimodal large language models (MLLMs), mainly due to their lack of sports diversity and task complexity. Specifically, the current domain of sports video analysis has been a rich and multifaceted domain that covers several visual tasks, ranging from re-id (Comandur, 2022), OCR (Solberg et al., 2024), temporal localization (Liu et al., 2022), fine-grained action classification (Liu et al., 2024d), motion understanding (Feng et al., 2024), etc. These tasks often focus on narrow tasks or evaluate only specific aspects of the content, without considering the full spectrum of multimodal information. For instance, in tasks such as sports video captioning, the focus is typically limited to key event captions, such as describing specific actions performed by athletes, while neglecting other critical aspects, including emotional context and background.

Sports commentary generation, unlike previous approaches, provides comprehensive coverage of sports video information by capturing key events, tactical analysis, emotional expressions, and background information. To this end, we propose the fine-grained CommentarySet, a novel and comprehensive dataset entailing a new commentary task designed to better utilize the diverse information embedded in sports videos. As demonstrated by Fig. 1, CommentarySet consists of 5,775 high-quality sports video clips spanning six diverse sports including athletics, basketball, soccer, gym, table tennis, and tennis, each meticulously annotated with professional English commentary. Furthermore, we introduce a novel six-dimensional commentary with human annotation and a GPT-based evaluation method grounded in our six-dimensional metric as shown in Fig. 1. The dataset and evaluation methods account for the dynamic and emotionally resonant nature of sports commentary, which requires not only accurate event descriptions but also a deep integration of contextual information. In our Extensive experiments also demonstrate that most multi-modal video models fail to perform successful sports commentary, which remains a challenging domain.

Our main contributions are summarized as follows:

- We introduce CommentarySet, the first dataset featuring professionally annotated sports video commentary across multiple disciplines, encompassing six sports and over ten international competitions.
- We propose a meticulously designed six-dimensional metric, enabling robust evaluation of advanced video language models. These account for the dynamic and emotional nature of sports commentary, which requires a deep integration of contextual information.
- We evaluate the in-domain and out-of-domain performance of several open-source pre-trained Video LLMs on CommentarySet. Our six-dimensional metric demonstrates significant limitations in existing models' capabilities in performing sports commentary generation in both cases.

2 Related Works

2.1 Benchmarking Video LLMs

Recently, the emergence of numerous Video LLMs (Lin et al., 2024; Liu et al., 2024a; Lin et al., 2023; Chen et al., 2024; Zhang et al., 2024) have been proven to be effective in solving various downstream tasks. Consequently, various benchmarks, based on different evaluation dimensions, have been introduced to evaluate their performance. Previous works like Perception Test (Pătrăucean et al., 2023), EgoSchema (Mangalam et al., 2023), and TempCompass (Liu et al., 2024c) have focused on specific aspects of the models' capabilities, such as spatial understanding and

temporal comprehension. More recent works, including Video-Bench (Ning et al., 2023b), Video-MME (Fu et al., 2024), MMBench-Video (Fang et al., 2024), MVBench (Li et al., 2024) and LVBench (Wang et al., 2024), evaluate models by defining task-specific hierarchical capability taxonomy of the model and adopts more diverse and longer videos.

2.2 Sports Videos for Video LLMs

2.2.1 Sports Video Datasets

Sports videos for video captioning models represent a branch within video understanding tasks, with numerous datasets and benchmarks proposed for various sports and tasks. Datasets such as TenniSet (Faulkner and Dick, 2017), FineGym (Shao et al., 2020), P2A (Bian et al., 2024), and SoccerNet (Deliège et al., 2021) provide detailed labels for technical actions and events in the games. Specifically, datasets like TenniSet, FineGym, and P2A contain a large number of standardized labels about on-field events or the description of on-field player actions. Another type of research focuses on interactions between athletes, exemplified by SportsHHI (Wu et al., 2024), which focuses on athletes' location and interaction classification tasks for sports videos. Although these datasets provide data for sports video recognition tasks of Video LLMs, there is still a lack of commentary-based sports videos, which is exactly what our newly proposed dataset addresses.

2.2.2 Video Captioning on Sport Videos

Sports datasets like SoccerNet (Deliège et al., 2021) and SportsHHI (Wu et al., 2024) have evaluated models' abilities in event understanding or interaction recognition tasks within sports videos. These captioning tasks assess models from a textual perspective such as feature or N-gram similarity. Despite advancements in datasets and benchmarks for technical actions and interactions, there remains a need for a comprehensive sports dataset and benchmark for commentary-based captioning - whose particularity has been mentioned above. That's why we've come up with a new commentary benchmark to measure this ability.

3 Evaluation & Metric

In this section, we first define the task of sports commentary, highlighting its distinctions from conventional video captioning tasks. Subsequently, we present our novel six-dimensional metric tailored for this task, compared with traditional metrics and the vanilla GPT-based method. Finally, we provide an in-depth description of our GPT-based evaluation method, grounded in the six-dimensional metric.

3.1 Task Definition

Sports commentary provides descriptions, analysis, and insights to enhance the viewing experience with detailed information, tactical breakdowns, and highlights. The task involves providing a video input and a prompt to the model. The demanding output is the corresponding commentary.

Unlike traditional video captioning, which only focuses on describing visual content, sports commentary requires not only accurate event descriptions but also timely, emotionally resonant narration that aligns with the game's pace and context. For example, when game intensity is relatively low, commentators may provide tactical analysis or details of the players or the teams, while remarkable plays might trigger emotionally charged reactions like "What a goal!" and "Amazing!" Fig. 3 illustrates the difference between caption and commentary.

3.2 Six-Dimensional Metric

Having established the definition of our task and distinguished it from video captioning, we now explain the rationale for our novel metric. We first compare it with traditional captioning metrics to demonstrate its necessity, followed by a comprehensive introduction to the six-dimensional metric.

3.2.1 Why Propose a New Metric

Existing benchmarks for Video LLMs mostly deploy metrics designed for either QA or multi-QA tasks, which are unsuitable for our task. Meanwhile, the metrics currently used for video captioning tasks fail to fully accommodate our specific requirements.

To be more specific, traditional metrics (e.g. **BLEU** (Papineni et al., 2002), **ROUGE_L** (Lin, 2004), **METEOR** (Banerjee and Lavie, 2005), **CIDEr** (Vedantam et al., 2014) and **SPICE** (Anderson et al., 2016)) primarily evaluate the lexical or structural similarity between generated texts and reference texts, focusing on aspects like n-gram overlap, sequence matching, and semantic accuracy. However, they do not capture the model's ability to understand fine-grained professional de-

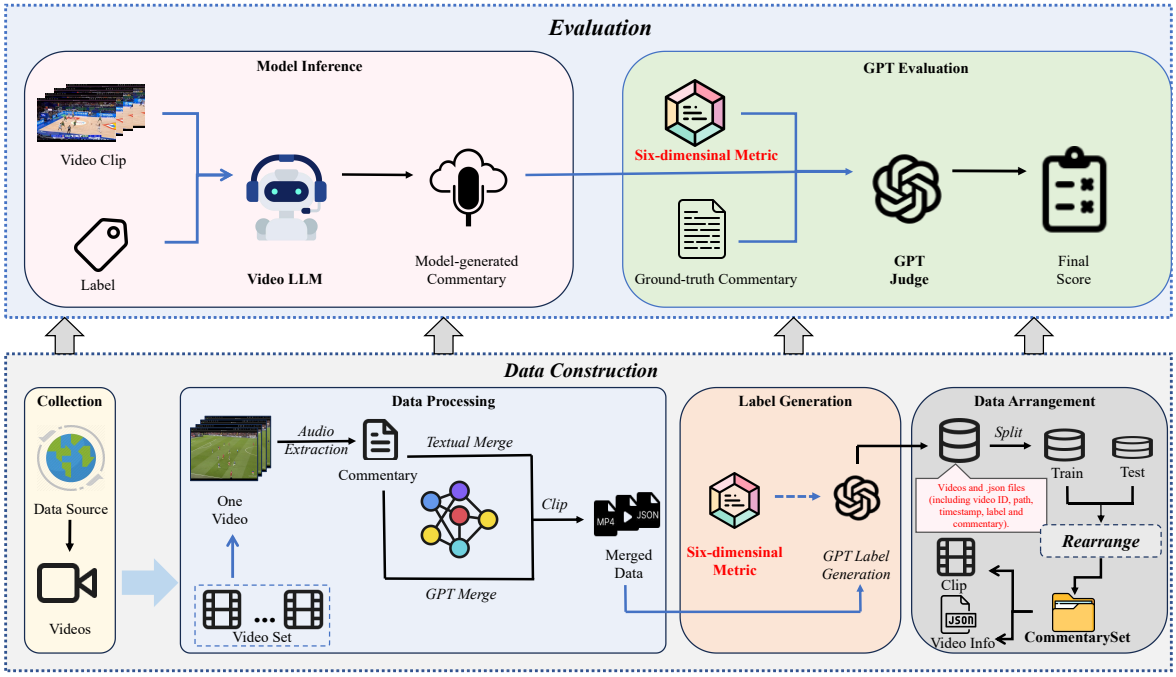

Figure 2: **The overall framework of SCBench**. SCBench consists of two parts: the construction of CommentarySet and our GPT-based evaluation.

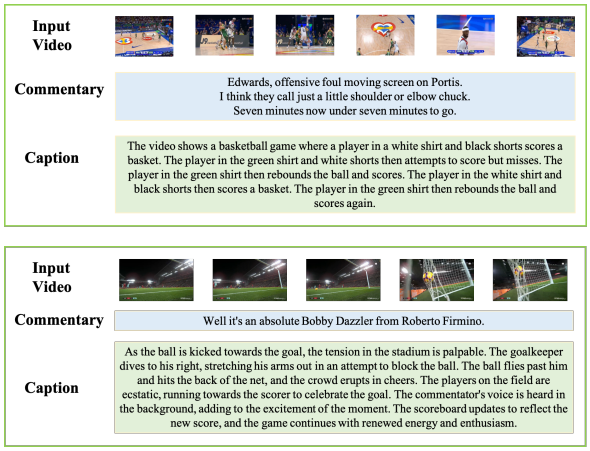

Figure 3: **Distinctions between Caption and Commentary.** Although captioning provides precise visual descriptions, commentary delivers dynamic, context-rich, and emotionally resonant narration.

tails, temporal dynamics, or human emotions in commentary. Apart from that, compared to the vanilla GPT approach, which directly evaluates the generated commentary based on ground truth, our method not only provides a more detailed and fine-grained scoring standard for GPT but also allows for a more intuitive demonstration of the model's capabilities across different dimensions.
Detailed experimental demonstrations of comparison are included in the supplementary material.

3.2.2 Six-Dimensional Metric

To address the limitations of traditional metrics and better evaluate the multifaceted nature of sports commentary, we propose our novel **Six-Dimensional Metric**, which includes the following aspects:

1. **Key Events Caption:** This dimension assesses the model's ability to accurately detect and describe significant events, such as scores, turnovers, fouls, and substitutions, reflecting its proficiency in real-time event detection and textual articulation.
2. **Technical Detail Analysis:** It evaluates models' capacity to analyze and explain intricate player actions, such as passing and shooting techniques, reflecting its fine-grained visual understanding capability and contextual knowledge integration.
3. **Background Information Interpretation:** This dimension measures the model's ability to integrate visual and textual information about players' and teams' histories, characteristics, and performance, which demonstrates its contextual knowledge integration and deep understanding of visual information(e.g. game intensity and event frequency).
4. **Tactical Analysis:** It assesses the model's skill in interpreting and explaining team strategies, formations, and in-game tactical adjustments by analyzing visual inputs and conveying these insights through commentary, highlighting its temporal understanding of game dynamics.
5. **Match Situation Interpretation:** This dimension evaluates the model's ability to interpret the current state of the match, such as the score and momentum, through visual cues and predict its likely

Table 1: Commentary statistics of our proposed dataset across different sports domains.

Domain	Clip Num	Avg Duration	Commentary Length	Commentary Freq	Source
Total	5,775	14.5s	30.14	45.73	-
Athletics	1,383	18.38s	44.77	30.44	IAAF Gold Label Road Races, IAAF Diamond League
Basketball	1,613	12.06s	23.36	26.18	2023 FIBA Basketball World Cup
Soccer	696	10.26s	19.79	70.51	English Premier League
Gym	1,153	16.49s	34.68	67.61	2015/2016 World Gymnastics Championship, etc.
Table Tennis	719	9.96s	20.63	31.7	2019-2022 World Cups
Tennis	211	26.27s	27.86	141.92	2012 London Olympics

progression, providing dynamic insights.

6. **Emotional Expression:** It measures the model's ability to detect emotional elements from visual data, such as player reactions and crowd energy, and effectively express these emotions in text to engage the audience emotionally.

3.3 GPT-based Evaluation

Built upon our six-dimensional metric, we propose a novel GPT-based evaluation method tailored for sports video commentary, as shown in Fig. 2. Here, we provide a detailed overview of evaluation processes.

The evaluation process could be divided into two stages: Model Inference and GPT Evaluation. In the first stage, we employ a three-tier prompt structure of **task definition - corresponding six-dimensional metric definition - task requirement**: For each video clip, we first provide the model with the definition of tour task, followed by the definition of the six-dimensional metric label for that clip. Finally, we input the task requirement into the model. As for **GPT Evaluation**, we input both the ground truth and model-generated commentary into our GPT judge, instructing it to evaluate the generated commentary based on the ground truth and the labels among the six dimensions. The evaluation yields a score from 0 to 10, with 0 indicating very poor quality and 10 representing perfection. Detailed prompts for these two steps are in supplementary material.

4 Dataset

Our goal is to advance research in the field of sports video commentary by introducing a challenging benchmark with professional commentary annotations. To this end, we have developed the fine-grained **CommentarySet**, which contains 5775 high-quality(1080P) sports video clips. Each video is well annotated, and the overall dataset structure will be discussed in the next section.

4.1 Dataset Construction

Data Structure To better evaluate sports video commentary, we create a novel dataset called CommentarySet for SCBench, containing a total of 5,775 clips, with 4,908 in the training set and 867 in the test set, the comparison with other datasets is shown in Tab. 2. Part of the clips is selected from FineGym (Shao et al., 2020) and TenniSet (Faulkner and Dick, 2017). Specifically, CommentarySet includes clips from six types of sports, along with their corresponding timestamps in the original videos. The necessary information on the six sports we selected is shown in Tab. 1. Moreover, each clip is accompanied by carefully selected English commentary text, which can serve as a ground truth label in sports commentary evaluation. To implement the evaluation metrics proposed in this paper, we also provide each clip with precise six-dimensional labels.

Table 2: Comparison of different datasets across different sports domains. **Avg Duration** is the average duration of all clips in the sport

Dataset	Domain	Clip Numbers	Avg Duration
TennisSet	Tennis	3,568	1.08s
FineGym	Gymnastic	4,885	8s
FSN	Basketball	2,000	-
SoccerNet-caption	Soccer	942	238s
CommentarySet (ours)	Multiple sports	5,775	14.5s

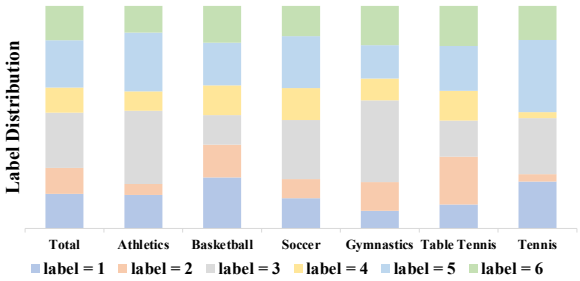

Figure 4: Distribution of the labels

Dataset Creation Process To create CommentarySet and ensure the reliability, professionalism, and accuracy of the video resources and commentary, we follow the pipeline below for dataset creation,

as shown in Fig. 2:

1. Collection: We collect a large number of high-definition sports videos with original English commentary from various sports and events available online and extracted the English commentary from the original audio, preserving their corresponding timestamps.

2. Merging: The commentary sentences obtained from the Collection process are fragmented. So, we add a process to merge them. To be specific, let the previous and current commentary be Com_1 and Com_2 respectively, and let their time stamps be $[\text{b}_1, \text{e}_1]$ and $[\text{b}_2, \text{e}_2]$. The merged commentary and its time stamp can be obtained using the following formula:

$$\text{New Com} = \text{Com}_1 + \text{Com}_2 \tag{1}$$

$$\text{New Time Stamp} = [b_1, e_2] \tag{2}$$

Regarding the determination of whether two commentaries should be merged, we conduct the following two-stage decision process:

Stage 1 Textual Merging: We encode the sentences first and then apply cosine similarity to merge commentary with high textual similarity such as repetitive cheers or consecutive sentences describing the same content. To be specific, the similarity (Sim) is calculated using the cosine similarity. After obtaining Sim, we simply decide whether to merge according to the following formula:

$$Action = \begin{cases} \text{Merge} & \text{if Sim} > \text{threshold} \\ \text{Separate} & \text{else} \end{cases} \tag{3}$$

where the threshold is set to 0.7 as our experimental choice.

Stage 2 GPT Merging: Some segments with continuous semantic meaning but no textual overlap remained. We use GPT-4o-mini to judge whether to merge consecutive commentary segments, the prompt would be shown in supplementary material. This is applied to segments where there exists a logical connection and describes the same topic.

3. Slicing: Before slicing, we manually refine the timestamps. The full sports match videos are sliced according to the timestamps, generating clip-commentary pairs.

4. Six-Dimensional Label Generation: Finally, we use GPT-4o-mini to classify the semantic content of the commentary. Each commentary sentence is categorized into one or more of the six dimensions mentioned in the metric section.

5. Human Refinement: After the aforementioned automated collection and cleaning procedures, we performed manual refinement on sampled data to ensure quality and consistency, ultimately organizing the processed results into our CommentarySet.

4.2 Statistic Analysis & Comparison

This section will primarily present specific statistical information about CommentarySet. As shown in Tab. 1, We will demonstrate that CommentarySet is a more suitable dataset for our proposed SCBench compared to other existing sports video captioning datasets, offering greater overall quality.

Diversity The six sports included in CommentarySet are carefully selected to ensure strong diversity. As shown in Tab. 1, the six selected sports exhibit significant diversity in terms of match speed, player density, commentary length, and commentary frequency. For instance, the average duration of table tennis clips is the shortest at just **9.96s**, while athletics, gymnastics, and tennis have clip lengths **1.5** to **2** times longer, reflecting faster match paces and shorter commentary content. In terms of commentary frequency, basketball has the highest frequency, while soccer and tennis have relatively lower frequencies, correlating closely with the frequency of events occurring on the field.

Commentary Style Commentary style is another critical aspect. The content and focus of commentary vary across different sports, which motivate our proposal of the six-dimensional metric. Through the analysis of the distribution of the six labels across sports, as shown in Fig. 4, we have observed distinct commentary styles. For example: In basketball, key event captions account for 44.33%, indicating continuous on-field events; In gymnastics, 61.84% of the commentary includes Background Information Interpretation, as commentators frequently introduce each participating athlete. The diversity of sports leads to diverse commentary styles, which is quantitatively reflected in the distribution of labels across different sports.

5 Experiments

We conduct extended experiments which could be divided into two main parts. First, we evaluated the zero-shot commentary capabilities of multiple open-source Video LLMs on our SCBench. Sub-

Table 3: **Performance of Video LLMs.** Including the pre-trained models and the fine-tuned models. "ICL" indicates utilizing the ICL fine-tuning dataset while "CoT" indicates using the CoT training set.

Model Name	Model Params	Traditional Metric		Performance in each sport event (0-10)						Average (0-10)
		BLEU(0-100)	CIDEr(0-1)	Table Tennis	Basketball	Soccer	Gym	Tennis	Athletics	
Mini-InternVL-Chat-4B-v1.5	4B	0.44	**0.12**	4.98	4.42	4.18	4.36	4.54	4.20	4.40
Video-LLaVA	7B	0.67	0.08	2.65	2.56	2.33	2.97	2.17	2.25	2.55
LongVA	7B	0.39	0.1	4.80	3.59	3.59	3.79	3.83	3.82	3.85
LLaVA-NeXT-Video	7B	0.35	0.1	4.77	4.39	4.22	4.24	4.28	3.65	4.20
Chat-UniVi-1.5	7B	0.56	0.11	3.67	3.48	2.96	3.68	2.70	3.08	3.37
Kangaroo	8B	0.83	0.01	1.60	1.64	1.70	1.97	2.02	1.89	1.79
VILA	34B	0.66	0.09	3.83	3.42	3.24	3.78	3.00	2.92	3.40
InternVL-Chat-2	**40.1B**	0.32	0.11	**5.92**	**5.50**	**5.42**	**5.31**	**4.98**	**5.29**	**5.44**
Video-LLaVA-ICL	7B	2.11	0.06	2.76	2.70	2.16	3.01	2.98	2.60	2.70
Chat-UniVi-1.5-ICL	7B	**2.57**	0.06	3.66	3.57	3.03	3.79	2.81	3.04	3.42
Video-LLaVA-CoT	7B	1.51	0.05	2.40	3,00	2.61	2.77	3.64	2.67	2.77
Chat-UniVi-CoT	7B	1.44	0.06	2.44	3.78	3.47	3.92	3.42	3.62	3.57

sequently, we assessed the effects of ICL (Brown et al., 2020) and CoT (Wei et al., 2023) methods on fine-tuneing Video-LLaVA and Chat-UniVi. In this section, we first introduce our experiment settings, including model selection and hyperparameters. Then we present quantitative results for both zero-shot experiments and fine-tuned models. Finally, we present a detailed visualization, analysis, and case study of the experimental results across different models and fine-tuning methods.

5.1 Settings and Configurations

All the experiments are conducted on 8 A800 GPUs. We perform evaluation on eight open-source pre-trained Video LLMs, including VILA (Lin et al., 2024), Video-LLaVA (Lin et al., 2023), LongVA (Zhang et al., 2024), LLaVA-NeXT-Video (Liu et al., 2024a), Kangaroo (Liu et al., 2024b), Chat-UniVi-1.5 (Jin et al., 2024), InternVL-2 (Team), and Mini-InternVL-v1.5 (Gao et al., 2024), as well as two fine-tuned models, which are Video-LLaVA and Chat-UniVi-1.5.
As is illustrated in the section 3.3, the evaluation process is divided into two main stages: Model Inference and GPT Evaluation. In the Model Inference stage, we follow their official configurations and try to use more frames and more maximum output tokens for commentary generation. For GPT Evaluation, we utilize *GPT4o-mini* model as our judge, with the temperature set to 0.1. As for fine-tuning, we follow the official document for full-parameter tuning and set the learning rate marginally lower. Detailed settings and prompts are shown in the supplementary material.

5.2 Quantitative Results

Zero-shot Model Performance To comprehensively evaluate the performance on our SCBench, we conduct extensive experiments on the eight aforementioned pretrained Video LLMs, the result is shown in Tab. 3. Overall, InternVL-2, with **40.1B** parameters, perform the best, achieving a score of **5.44**, surpassing second-best, Mini-InternVL-Chat-v1.5, by **1.04**. Following this, LLaVA-NeXT-Video, LongVA and VILA score **4.20**, **3.85** and **3.40**, respectively. Our evaluation is based on a six-dimensional metric. So we calculate the models' performance in all six dimensions and create a six-dimensional radar chart, as shown in Fig. 5, showing the comprehensive performance of models. Additionally, as the models show varying strengths across different sports, we also create a six sports radar chart in Fig. 5 to show models' performance across different sports types. The strengths across different sports highlight the models' specific capabilities mentioned in the dataset section.
ICL & CoT Fine-tuning We propose two methods of fine-tuning and evaluate them with Video-LLaVA and Chat-UniVi on our SCBench:

1. **In-Context Learning (ICL) Fine-tuning**: In-Context Learning refers to the ability to perform specific tasks by leveraging contextual information embedded within the prompt. In our fine-tuning, We structure the fine-tuning QA setup similar to the prompt format used in Model Inference, instructing the model to directly generate commentary based on the video, task definition, and the perspectives of the ground truth metrics, intended to enhance the model's comprehension of the task and the metric perspectives and align the characteristics of the metric dimensions with the generated commentary.

2. **Chain-of-Thought (CoT) Fine-tuning**: Chain-of-Thought (CoT) is a prompting technique that enhances models' reasoning by breaking down complex problems into smaller steps. This ap-

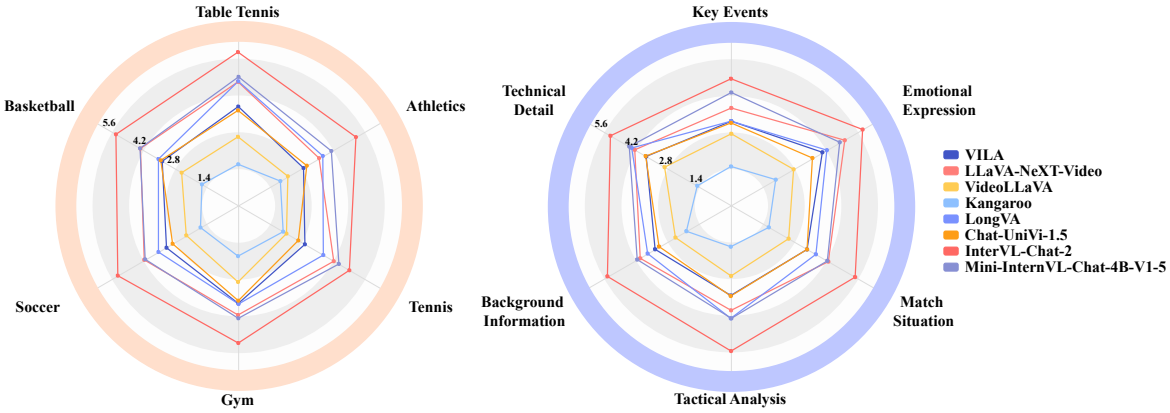

Figure 5: Model Performances on Six Sports (left) and Six-Dimensional (right) tasks

proach, which is more challenging than providing a specific metric dimension, aims to encourage the model to emulate human reasoning in commentary, enhancing coherence and conciseness.

The scores of fine-tuned models are demonstrated in Tab. 3. The results indicate that both fine-tuning methods enhance the model's performance, with CoT providing a more substantial improvement. Specifically, CoT increased performance by **8%** on Video-LLaVA and **6%** on Chat-UniVi, while ICL improved performance by **6%** and **1.5%**, respectively. However, although the model's performance has improved, the score doesn't increase as significantly as observed in other tasks (for instance, the best-performing fine-tuned model still does not surpass LongVA or Mini-InternVL). We'll illustrate this point in the next section.

5.3 Analysis

We perform analysis to further demonstrate the superiority of our metric over traditional metrics and discuss the possible factors that influence the zero-shot and fine-tuned models' performance.

5.3.1 Quantitative Comparison

In Sec. 3, we theoretically demonstrate the necessity of proposing a new six-dimensional metric for sports commentary tasks. We further conduct experiments using BLEU and CIDEr to evaluate different models, as shown in Tab. 3. The results indicate that traditional metrics yield low scores across all models (BLEU scores below 1, CIDEr scores around 0.1), making it challenging to distinguish model performance. In contrast, the six-dimensional metric offers a more nuanced evaluation, effectively and distinctively measuring the model's capabilities. For example, although InterVL-Chat-2 has the lowest BLEU score and a relatively low CIDEr score, our metric highlights its superior performance across all categories.

5.4 Zero-shot vs Fine-tune

Grounded on this, we can interpret many of our experimental results. For zero-shot models, we observed that InternVL-2 achieved significantly higher scores than other models. Upon examining its generated commentaries, we find that its outputs were notably longer, indicating that it captures more concise and accurate visual or background information. Apart from this, in the case of fine-tuning, the two methods we proposed primarily enable models to comprehend the stylistic elements of training data and the specific content of the metric. However, these methods do not enhance the models' ability to capture pivotal visual or background information, which explains the reason for the fine-tuned models' underperformance compared with certain zero-shot models. Therefore, we believe that establishing a keyword repository for the model to perform Retrieval-Augmented Generation (RAG) (Lewis et al., 2021) or employing methods such as VisCoT (Shao et al., 2024) to enhance the model's ability to capture visual information would be key strategies for further improving the model's commentary capabilities.

6 Conclusion

In this paper, we introduce a novel benchmark named **SCBench** for Video LLMs, focusing on the complex task of sports video commentary. We develop CommentarySet, a dataset comprising 5,775 annotated clips across six sports categories, and propose a six-dimensional metric tailored to evaluate the task. Our experiments reveal significant limitations in the ability of current models to generate effective sports commentary, providing a foundation for advancing Video LLMs' capabilities in complex visual understanding and text generation.

7 Limitations

The limitations of our work can be summarized as follows: First, our evaluation primarily relies on the closed-source large model GPT-4o-mini-2024-07-18, which may introduce potential unfairness and bias. Second, although our dataset encompasses six types of sports and is constructed through an automated data collection and cleaning pipeline, its diversity remains limited and primarily consists of English-language content, which may introduce linguistic and cultural bias. Third, despite experimenting with two fine-tuning strategies, the current models still exhibit significant deficiencies in capturing key visual events and contextual information, suggesting that future work should explore alternative fine-tuning or adaptation methods to further enhance model performance.

References

Peter Anderson, Basura Fernando, Mark Johnson, and Stephen Gould. 2016. Spice: Semantic propositional image caption evaluation. *Preprint*, arXiv:1607.08822.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, Ann Arbor, Michigan. Association for Computational Linguistics.

Jiang Bian, Xuhong Li, Tao Wang, Qingzhong Wang, Jun Huang, Chen Liu, Jun Zhao, Feixiang Lu, Dejing Dou, and Haoyi Xiong. 2024. P2anet: A dataset and benchmark for dense action detection from table tennis match broadcasting videos. *Preprint*, arXiv:2207.12730.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. *Preprint*, arXiv:2005.14165.

Zhe Chen, Jiannan Wu, Wenhai Wang, Weijie Su, Guo Chen, Sen Xing, Muyan Zhong, Qinglong Zhang, Xizhou Zhu, Lewei Lu, Bin Li, Ping Luo, Tong Lu, Yu Qiao, and Jifeng Dai. 2024. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. *Preprint*, arXiv:2312.14238.

Bharath Comandur. 2022. Sports re-id: Improving re-identification of players in broadcast videos of team sports. *Preprint*, arXiv:2206.02373.

Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. 2021. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. *Preprint*, arXiv:2011.13367.

Xinyu Fang, Kangrui Mao, Haodong Duan, Xiangyu Zhao, Yining Li, Dahua Lin, and Kai Chen. 2024. Mmbench-video: A long-form multi-shot benchmark for holistic video understanding. *Preprint*, arXiv:2406.14515.

Hayden Faulkner and Anthony R. Dick. 2017. Tenniset: A dataset for dense fine-grained event recognition, localisation and description. *2017 International Conference on Digital Image Computing: Techniques and Applications (DICTA)*, pages 1–8.

Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J Black. 2024. Chatpose: Chatting about 3d human pose. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2093–2103.

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

Zhangwei Gao, Zhe Chen, Erfei Cui, Yiming Ren, Weiyun Wang, Jinguo Zhu, Hao Tian, Shenglong Ye, Junjun He, Xizhou Zhu, Lewei Lu, Tong Lu, Yu Qiao, Jifeng Dai, and Wenhai Wang. 2024. Mini-internvl: A flexible-transfer pocket multimodal model with 5% parameters and 90% performance. *Preprint*, arXiv:2410.16261.

Peng Jin, Ryuichi Takanobu, Wancai Zhang, Xiaochun Cao, and Li Yuan. 2024. Chat-univi: Unified visual representation empowers large language models with image and video understanding. *Preprint*, arXiv:2311.08046.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. 2024. Mvbench: A comprehensive multi-modal video understanding benchmark. *Preprint*, arXiv:2311.17005.

Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. 2023. Video-llava: Learning united visual representation by alignment before projection. *Preprint*, arXiv:2311.10122.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Ji Lin, Hongxu Yin, Wei Ping, Yao Lu, Pavlo Molchanov, Andrew Tao, Huizi Mao, Jan Kautz, Mohammad Shoeybi, and Song Han. 2024. Vila: On pre-training for visual language models. *Preprint*, arXiv:2312.07533.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024a. Llava-next: Improved reasoning, ocr, and world knowledge.

Jiajun Liu, Yibing Wang, Hanghang Ma, Xiaoping Wu, Xiaoqi Ma, Xiaoming Wei, Jianbin Jiao, Enhua Wu, and Jie Hu. 2024b. Kangaroo: A powerful video-language model supporting long-context video input. *Preprint*, arXiv:2408.15542.

Yi Liu, Limin Wang, Yali Wang, Xiao Ma, and Yu Qiao. 2022. Fineaction: A fine-grained video dataset for temporal action localization. *IEEE transactions on image processing*, 31:6937–6950.

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

Zhendong Liu, Haifeng Xia, Tong Guo, Libo Sun, Ming Shao, and Siyu Xia. 2024d. Cross-block fine-grained semantic cascade for skeleton-based sports action recognition. *Preprint*, arXiv:2404.19383.

Karttikeya Mangalam, Raiymbek Akshulakov, and Jitendra Malik. 2023. Egoschema: A diagnostic benchmark for very long-form video language understanding. *Preprint*, arXiv:2308.09126.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2023a. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *arXiv preprint arXiv:2311.16103*.

Munan Ning, Bin Zhu, Yujia Xie, Bin Lin, Jiaxi Cui, Lu Yuan, Dongdong Chen, and Li Yuan. 2023b. Video-bench: A comprehensive benchmark and toolkit for evaluating video-based large language models. *Preprint*, arXiv:2311.16103.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Viorica Pătrăucean, Lucas Smaira, Ankush Gupta, Adrià Recasens Continente, Larisa Markeeva, Dylan Banarse, Skanda Koppula, Joseph Heyward, Mateusz Malinowski, Yi Yang, Carl Doersch, Tatiana Matejovicova, Yury Sulsky, Antoine Miech, Alex Frechette, Hanna Klimczak, Raphael Koster, Junlin Zhang, Stephanie Winkler, and 5 others. 2023. Perception test: A diagnostic benchmark for multimodal video models. *Preprint*, arXiv:2305.13786.

Kate Sanders and Benjamin Van Durme. 2024. A survey of video datasets for grounded event understanding. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7314–7327.

Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. 2020. Finegym: A hierarchical video dataset for fine-grained action understanding. *Preprint*, arXiv:2004.06704.

Hao Shao, Shengju Qian, Han Xiao, Guanglu Song, Zhuofan Zong, Letian Wang, Yu Liu, and Hongsheng Li. 2024. Visual cot: Advancing multi-modal language models with a comprehensive dataset and benchmark for chain-of-thought reasoning. *Preprint*, arXiv:2403.16999.

Håkon Maric Solberg, Mehdi Houshmand Sarkhoosh, Sushant Gautam, Saeed Shafiee Sabet, Pål Halvorsen, and Cise Midoglu. 2024. Playertv: Advanced player tracking and identification for automatic soccer highlight clips. *Preprint*, arXiv:2407.16076.

OpenGVLab Team. Internvl2: Better than the best—expanding performance boundaries of open-source multimodal models with the progressive scaling strategy. Accessed: 2024-11-14.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2014. Cider: Consensus-based image description evaluation. *CoRR*, abs/1411.5726.

Weihan Wang, Zehai He, Wenyi Hong, Yean Cheng, Xiaohan Zhang, Ji Qi, Shiyu Huang, Bin Xu, Yuxiao Dong, Ming Ding, and Jie Tang. 2024. Lvbench: An extreme long video understanding benchmark. *Preprint*, arXiv:2406.08035.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. Chain-of-thought prompting elicits reasoning in large language models. *Preprint*, arXiv:2201.11903.

Tao Wu, Runyu He, Gangshan Wu, and Limin Wang. 2024. Sportshhi: A dataset for human-human interaction detection in sports videos. *Preprint*, arXiv:2404.04565.

Peiyuan Zhang, Kaichen Zhang, Bo Li, Guangtao Zeng, Jingkang Yang, Yuanhan Zhang, Ziyue Wang, Haoran Tan, Chunyuan Li, and Ziwei Liu. 2024. Long context transfer from language to vision. *Preprint*, arXiv:2406.16852.

Pengyuan Zhou, Lin Wang, Zhi Liu, Yanbin Hao, Pan Hui, Sasu Tarkoma, and Jussi Kangasharju. 2024. A survey on generative ai and llm for video generation, understanding, and streaming. *arXiv preprint arXiv:2404.16038*.