WHEN, WHERE AND WHAT TO COMMENT: A CLUE-GROUNDED BENCHMARK FOR SPORTS VIDEO COMMENTARY

Anonymous authors
Paper under double-blind review

ABSTRACT

Real-world sports commentary requires more than generating descriptions from video clips. A competent commentator must continuously determine when an event deserves commentary, identify where the supporting visual evidence occurs, and decide what to say based on the event and its competitive significance. Existing benchmarks largely reduce this process to text generation from pre-segmented clips, implicitly specifying commentary timing and weakening the need for fine-grained visual grounding. We introduce CG-SC-Bench, a clue-grounded benchmark for sports commentary that evaluates these three capabilities in a unified framework. Each commentary instance is associated with explicit visual clues, enabling the benchmark to assess whether a model identifies commentary-worthy moments (When), localizes the evidence supporting the commentary (Where), and generates factually consistent commentary conditioned on the observed clues (What). CG-SC-Bench covers 8 major sports categories and 27 subcategories, comprising 2,684 commentary segments with fine-grained clue annotations. Based on these annotations, we construct three complementary evaluation tasks for commentary timing, visual clue localization, and clue-conditioned commentary generation. We evaluate a broad range of closed- and open-source vision-language models on CG-SC-Bench. The results reveal substantial limitations of current models in temporal decision-making, fine-grained visual grounding, and evidence-grounded factual generation, showing that fluent sports descriptions do not necessarily translate into reliable continuous commentary. CG-SC-Bench provides a diagnostic evaluation framework for developing sports video models that can comment at the right time, ground their statements in the right evidence and say the right thing.

1 INTRODUCTION

Multimodal large language models (MLLMs) have advanced rapidly in video understanding, showing increasing ability to perceive and reason about spatiotemporal events. Sports commentary, however, requires more than recognizing events. a model must decide what to describe, when to describe it, and why it matters in the evolving context of a competition. Existing benchmarks have examined complementary aspects of this task, including timestamped soccer commentary (Mkhallati et al., 2023), knowledge-grounded soccer captioning (Qi et al., 2023), commentary quality across diverse sports (Ge et al., 2024), and fine-grained commentary generation in continuous basketball streams (Wu et al., 2026).

Early benchmarks mainly formulate sports commentary as generation from pre-segmented clips or temporally anchored events. GOAL (Qi et al., 2023) studies knowledge-grounded soccer commentary, SoccerNet-Caption (Mkhallati et al., 2023) links commentaries to event timestamps, and MatchTime (Rao et al., 2024) improves video–commentary alignment. SCBench (Ge et al., 2024) extends evaluation to six sports but still relies on short clips. Recent works further study category-conditioned generation and narration rhythm in boxing (Wang et al., 2026), and causal fine-grained commentary in continuous basketball streams (Wu et al., 2026). However, they do not jointly evaluate visual grounding, causal decisions of when and what to comment on, and factually consistent

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

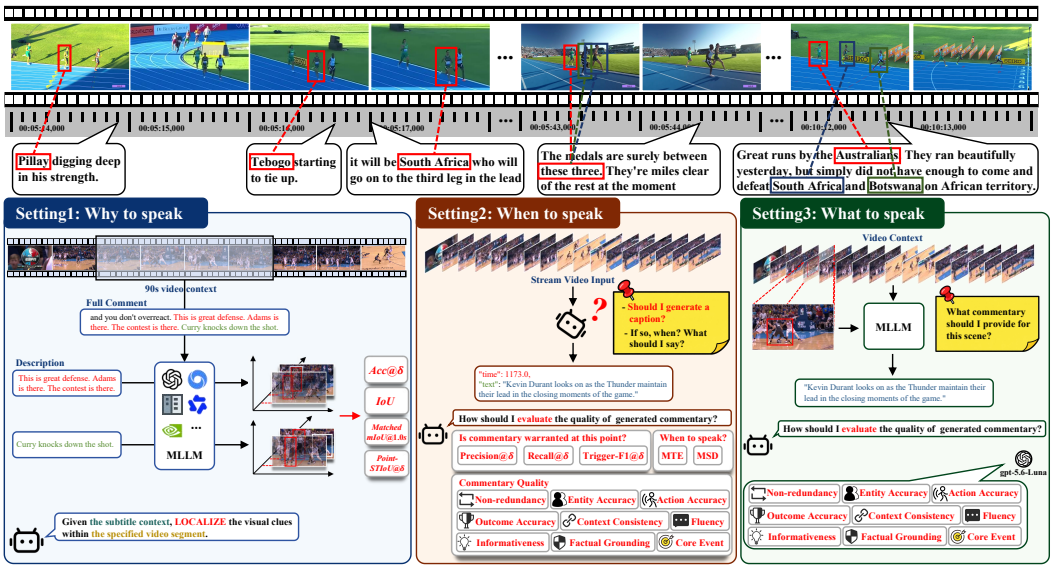


Figure 1: *Top*: Examples of visual clue annotations in CG-SC-Bench. Each commentary segment is linked to the relevant timestamps and bounding boxes, making its supporting visual evidence explicit and verifiable. *Bottom*: Three evaluation settings: commentary clue grounding, streaming commentary decision-making, and visual clue conditioned commentary generation.

generation conditioned on explicit visual regions. A comprehensive multi-sport benchmark integrating these capabilities is still lacking.

To make up this gap, we introduce CG-SC-Bench, illustrated in Figure 1, a clue-grounded comprehensive benchmark designed to evaluate video understanding for sports commentary. In contrast to existing benchmarks that primarily assess the quality of generated commentary, CG-SC-Bench comprises three complementary tasks. 1) *Commentary-to-spatiotemporal grounding* requires the model to identify the temporal interval and spatial regions of the relevant entities corresponding to a given commentary sentence. 2) *Causal streaming commentary decision* requires the model to determine whether commentary is needed, when it should be produced, and what should be described using only the video frames observed so far. 3) *Box-conditioned commentary generation* requires the model to generate commentary based on a video clip and target entities specified by bounding boxes, thereby evaluating controllability and factual consistency. Together, these tasks provide a unified evaluation of the grounding, decision-making, and generation capabilities required for realistic sports commentary.

CG-SC-Bench contains more than 12 hours continuous competition footage across eight sports categories and 27 subcategories. It includes 2,684 commentary segments, of which 2,168 have detailed visual evidence annotations. By preserving competition context beyond isolated highlights, the benchmark supports joint evaluation of visual grounding, causal decision-making, and conditioned generation.

Our experiments reveal that current MLLMs still struggle with fine-grained visual grounding, commentary timing, and factual generation. Even when models produce fluent and plausible commentary, they often fail to identify the correct evidence, trigger at the right moment, or maintain factual consistency with the observed event.

2 Related Work

Sports Video Understanding. Early sports video benchmarks focused on specialized perception tasks, including fine-grained action recognition with FineGym (Shao et al., 2020) and spatiotemporal action detection with MultiSports (Li et al., 2021). Subsequent datasets addressed action quality

assessment with FineDiving (Xu et al., 2022), multi-object tracking with SportsMOT (Cui et al., 2023), and human-human interaction detection with SportsHHI (Wu et al., 2024).

More recent work has expanded toward multimodal sports reasoning. SPORTU (Xia et al., 2025) evaluates sports knowledge and video question answering, while SportR (Xia et al., 2026) and RefereeBench (Xu et al., 2026) examine visual evidence, rules, and officiating decisions. DeepSport (Zou et al., 2025) develops multi-sport video reasoning through active frame selection. However, these advances do not provide a unified evaluation of how sports commentary aligns with fine-grained visual evidence in continuous, multi-sport videos.

Sports Commentary Benchmarks and Evaluation. Early work centered on soccer: SoccerNet-Caption (Mkhallati et al., 2023) introduced timestamped commentary captioning, GOAL (Qi et al., 2023) incorporated domain knowledge, and MatchTime (Rao et al., 2024) improved video–commentary alignment. BH-Commentary (Zhang et al., 2024) extended commentary generation to basketball highlights, while SCBench (Ge et al., 2024) evaluated generated commentary across multiple sports. Recent benchmarks examine commentary categories and narration rhythm in boxing with BoxComm (Wang et al., 2026), expert analytical commentary in tennis with TennisVL (Liu et al., 2026), and causal generation in continuous basketball streams with NBA_Streaming (Wu et al., 2026). Together, these works extend sports commentary research from timestamped event descriptions toward knowledge-rich and streaming generation.

Evaluation Frameworks for Sports Commentary. Early evaluations measured similarity to reference commentaries using BLEU (Papineni et al., 2002), METEOR (Banerjee & Lavie, 2005), ROUGE (Lin, 2004), and CIDEr (Vedantam et al., 2015). SoccerNet-Caption (Mkhallati et al., 2023) additionally matches predicted and reference commentaries within a temporal tolerance window before evaluating their text. To account for valid descriptions that differ in wording, SCBench (Ge et al., 2024) introduces the GPT-based, six-dimensional SCORES framework. BoxComm (Wang et al., 2026) combines semantic and LLM-based content evaluation with temporal IoU and KL divergence for commentary timing and rhythm, while NBA_Streaming (Wu et al., 2026) assesses localization, response reliability, factual grounding, and commentary quality under causal constraints. A unified evaluation of commentary timing, content, and grounding in corresponding visual entities across diverse sports remains lacking.

3 CG-SC-Bench: A Clue-Grounded Benchmark for Sports Commentary

3.1 Dataset Construction

We construct CG-SC-Bench in three stages: video collection, fine-grained alignment of commentary with visual evidence and iterative human review. We describe each stage below.

Video Collection. To preserve the progression and context of sports competitions, we collect publicly available videos primarily from official event channels and sports broadcasters. Each video covers a complete, temporally continuous competition segment, such as an entire quarter, race, bout or event stage, rather than an isolated highlight. We prioritize clear footage, intelligible commentary, and identifiable participants, retaining the original audio. Multiple rounds of human review remove duplicate, incomplete, poor-quality, or substantially misaligned videos.

Fine-grained Alignment of Commentary with Visual Evidence. We annotate commentary–visual evidence correspondence in two stages. First, an automatic speech recognition (ASR) model transcribes the commentary audio. Annotators verify the transcripts against the original audio, correct athlete names and sport-specific terms, and group the verified text into coherent commentary segments. Second, they identify visual cues referred to by each segment, including athletes, referee actions, and on-screen scoreboards, and annotate their times and bounding boxes. For streaming evaluation, annotations generally rely on footage available no later than the commentary. Commentary without directly identifiable visual cues is retained without forced visual alignment.

Iterative Human Review. The review covers four aspects: (1) *record integrity*: checking for missing fields and inconsistent timestamps; (2) *visual support*: verifying that cue descriptions match the annotated regions and can be confirmed in the video; (3) *temporal validity*, checking whether streaming annotations rely on future frames or confuse broadcast replays with live action; (4) *ref-*

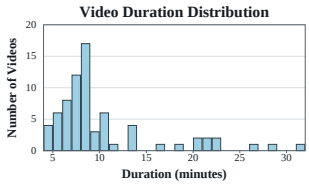


Figure 2: Video duration distribution, showing the number of videos in each duration interval.

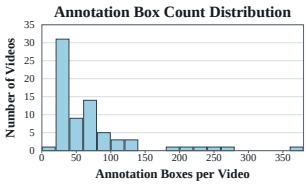


Figure 3: Distribution of visual evidence bounding boxes, showing a long-tail pattern.

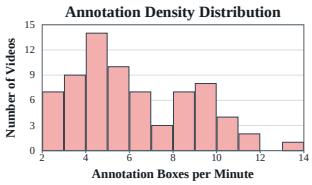


Figure 4: Annotation density distribution, reflecting variation in visual grounding difficulty.



Figure 5: Video duration distribution. The inner ring shows broad competition categories, while the outer shows specific sports.



Figure 6: Word cloud of sports and events, showing the variety of CG-SC-Bench,

erence clarity: flagging cases where an athlete or other visual referent cannot be identified reliably. Revised records are reviewed again, while unresolved visual links are excluded from the grounded subset.

3.2 Dataset Statistics & Comparisons

We provide a comprehensive overview of dataset statistics, including video metadata, commentary–visual evidence annotations, qualitative analysis, and comparisons with previous works.

3.2.1 Dataset Statistics

Video Metadata. CG-SC-Bench comprises 72 videos, each covering a complete, temporally continuous unit of competition. As shown in Figure 2, video durations range from 4 to 32 minutes, with more videos concentrated below 10 minutes. This duration profile helps prevent long-range video memory demands from dominating the evaluation. All videos were manually screened for visual quality and commentary clarity. Notably, the selected sports vary in competition pace, participant structure and scoring systems, providing diverse competitive settings. As shown in Figure 5, the dataset covers 8 major sports categories and 27 subcategories, including races, field events, head-to-head contests, team competitions, and judged performances. Further statistics are provided in the supplementary material.

Commentary-Visual Evidence Annotations. The dataset contains 2,684 timestamped commentary segments, of which 2,168 (80.8%) are grounded in visual evidence. These grounded segments are associated with 4,943 bounding boxes identifying relevant athletes, officials, scoreboards, and other visible cues. This extensive object-level grounding provides fine-grained supervision for evaluating

Table 1: Comparison of benchmarks across key aspects: number of sports (#Sport), number of videos (#Video), average duration(Dur.), number of Visual Cues Linked to Commentary(#VC), open-ended commentary generation(OE), temporal grounding (TG), spatial grounding (SG), multi-object grounding(MO), visual-cue-guided generation (VCG), commentary–timestamp alignment (CTA) and commentary–visual evidence alignment (CVEA).

Benchmark	#Sport	Dur.	#VC	#Commentary	Continuous	OE	TG	SG	MO	Streaming	VCG	CTA	CVEA
Sports Video Understanding													
FineGym (Shao et al., 2020)	1	2.34 h	0	–	✓	×	✓	×	×	×	×	–	–
SPORTU (Xia et al., 2025)	7	–	0	–	×	×	×	×	×	×	×	–	–
Sports Commentary and Captioning													
NSVA (Wu et al., 2022)	1	9.5 s	0	44,649	×	✓	×	×	×	×	×	×	×
SoccerNet-Caption (Mkhallati et al., 2023)	1	1.52 h	0	36,894	✓	✓	✓	×	×	×	×	✓	×
GOAL (Qi et al., 2023)	1	10.3 s	0	22K	×	✓	×	×	×	×	×	△	×
MatchTime (Rao et al., 2024)	1	1.52 h	0	32,743	✓	✓	✓	×	×	×	×	✓	×
SCBench (Ge et al., 2024)	6	14.5 s	0	5,775	×	✓	×	×	×	×	×	✓	×
NBA-Identity (Xi et al., 2025)	1	3.3 s	✓	9,726	×	✓	×	×	×	×	✓	△	△
BoxComm (Wang et al., 2026)	1	10.5 min	0	52K+	✓	✓	✓	×	×	✓	△	✓	×
NBA_Streaming (Wu et al., 2026)	1	2.02 h	0	50.4K	✓	✓	✓	×	×	✓	△	✓	×
Spatio-temporal Video Grounding													
VidSTG (Zhang et al., 2020)	–	28.01 s	0	–	–	×	✓	✓	×	×	×	–	–
HC-STVG (Tang et al., 2021)	–	20 s	0	–	–	×	✓	✓	×	×	×	–	–
BOSTVG (Yao et al., 2026)	–	–	0	–	–	×	✓	✓	✓	×	×	–	–
CG-SC-Bench (Ours)	**8**	**10.3 min**	**4,943**	**2,684**	✓	✓	✓	✓	✓	✓	✓	✓	✓

whether models can accurately associate commentary with the relevant participants, objects, and contextual cues, rather than relying solely on coarse event semantics. Further details about the dataset are provided in the appendix.

Figures 3 and 4 show the distributions of annotation boxes per video and per minute, respectively. The per-video distribution exhibits a clear long-tail pattern: most videos contain a moderate number of visual evidence boxes, whereas a small subset contains substantially more boxes. Although raw annotation counts are naturally influenced by video duration, considerable variation persists in the per-minute density. This variation may reflect differences in sport-specific event rhythms, participant interactions, and commentary styles. Taken together, these distributions expose models to a broad range of grounding conditions, from sparsely to densely annotated videos.

3.2.2 Comparison with Previous Benchmarks

Table 1 compares CG-SC-Bench with benchmarks in three areas: sports video understanding, sports commentary generation, and spatiotemporal video grounding. Sports understanding benchmarks, such as FineGym (Shao et al., 2020) and SPORTU (Xia et al., 2025), evaluate fine-grained action recognition or sports-specific question answering, but do not align commentary with localized visual evidence. Commentary benchmarks, including SoccerNet-Caption (Mkhallati et al., 2023), GOAL (Qi et al., 2023), MatchTime (Rao et al., 2024), SCBench (Ge et al., 2024), NBA-Identity (Xi et al., 2025), BoxComm (Wang et al., 2026), and NBA_Streaming (Wu et al., 2026), cover domain-specific descriptions of pre-segmented clips or continuous competitions. Some provide temporal alignment or structured event annotations, but most do not link individual commentary segments to spatial visual evidence. NBA-Identity (Xi et al., 2025) provides key-player bounding boxes, although its alignment focuses on player identities in short basketball clips.

Spatiotemporal grounding benchmarks, such as VidSTG (Zhang et al., 2020), HC-STVG (Tang et al., 2021), and BOSTVG (Yao et al., 2026), evaluate temporal, spatial, or multi-object localization using general-domain language queries. They do not address commentary generation or causal decisions about when to comment. CG-SC-Bench brings these capabilities together across complete, continuous competition units spanning 8 sports categories and 27 disciplines. It contains 2,684 timestamped commentary segments, including 2,168 segments aligned with 4,943 visual evidence boxes. This design supports a unified evaluation of commentary generation, spatiotemporal grounding, multi-object localization, and causal streaming commentary.

3.3 EVALUATION

In this section, we introduce the three evaluation tasks in SC-CG-Bench, including commentary-to-spatiotemporal grounding evaluation

3.3.1 SETTING 1: COMMENTARY-TO-SPATIOTEMPORAL GROUNDING EVALUATION.

This task evaluates whether a model can localize the visual evidence for a given commentary description. Each sample provides the competition name, commentary context, and video frames sampled at 1 FPS with their original-video timestamps. Given a target description d, the model predicts one or more timestamp–bounding-box pairs, $P_d = \{(\hat{t}_i, \hat{B}_i)\}$. Ground-truth pairs outside the evaluated segment $[s, e]$ are removed from $G_d = \{(t_j, B_j)\}$, yielding G'_d. Descriptions with no remaining ground-truth box are excluded.

Because the annotations mark representative frames rather than continuous object tubes, we use point-based localization metrics. For each temporal tolerance δ, predicted and ground-truth boxes are matched one-to-one using the Hungarian algorithm, first maximizing the number of pairs within δ and then their total spatial IoU. Let $\mathcal{M}_\delta$ be the resulting valid pairs and $N = \sum_d \max(|G'_d|, |P_d|)$. We compute

$$\text{Acc}\,@\delta = \frac{|\mathcal{M}_\delta|}{N}, \qquad \text{Point-STIoU}\,@\delta = \frac{\sum_{(i,j)\in\mathcal{M}_\delta} \text{IoU}(\hat{B}_i, B_j)}{N}, \quad \delta \in \{0.5, 1.0, 2.0\}\,\text{s}. \tag{1}$$

Unmatched boxes contribute zero to both metrics. We also report Mean Temporal Error (MTE) over matched pairs and $\text{Matched-mIoU}\,@1.0\text{s}$, the mean spatial IoU of pairs matched within one second.

3.3.2 SETTING 2: CAUSAL STREAMING COMMENTARY DECISION.

This task evaluates whether a model can decide when and what to comment on in a continuous sports video using only frames observed so far. The model also receives its previously generated commentaries. Let $G = \{(s_j, e_j, c_j, B_j)\}_{j=1}^{N_g}$ denote the reference commentaries, including their time intervals, texts, and visual-evidence annotations, and let $P = \{(\hat{t}_i, \hat{c}_i)\}_{i=1}^{N_p}$ denote the model's outputs. We use s_j as the reference trigger time. Predictions within δ seconds of a reference trigger are matched one-to-one, first maximizing the number of matches and then minimizing total temporal error. From the matched set $\mathcal{M}_\delta$, we report trigger precision, recall, F1, Mean Temporal Error (MTE), and Mean Signed Delay (MSD).

Commentary quality. For each matched pair, GPT-6-Sol assess the generated commentary against the reference, prior model-generated commentaries, and annotated visual evidence. They judge event, entity, action, and outcome accuracy; context consistency, relevance, informativeness, factual grounding, non-redundancy, clarity, and visual-evidence coverage. Each criterion receives a weighted `MET`, `UNMET`, or `NA` verdict, resolved by criterion-level majority voting. With verdict v_{ik} and weight w_k, we define

$$R_i = \frac{\sum_k w_k \mathbf{1}[v_{ik} = \text{MET}]}{\sum_k w_k \mathbf{1}[v_{ik} \in \{\text{MET}, \text{UNMET}\}]}, \qquad \text{Rubric-F1}@\delta = \frac{2\sum_{(i,j)\in\mathcal{M}_\delta} R_i}{N_p + N_g}. \tag{2}$$

Criteria marked `NA` are excluded. Rubric-F1 thus accounts for commentary quality as well as missed and redundant commentaries.

Causal consistency. For each correctly expressed visual-evidence claim k in prediction i, let τ_{ik} be the earliest time its evidence appears. A claim is causally valid if $\tau_{ik} \le \hat{t}_i + \epsilon$, where $\epsilon = 1.0\,\text{s}$. Let C_i be the mean validity of the assessable claims in commentary i. We report

$$\text{Causal-Rubric-F1}@\delta = \frac{2\sum_{(i,j)\in\mathcal{M}_\delta} R_i C_i}{N_p + N_g}. \tag{3}$$

We also report the Causal Consistency Rate (CCR) and Future Leakage Rate ($\text{FLR} = 1 - \text{CCR}$). All temporal and joint metrics use $\delta \in \{1, 2, 5\}\,\text{s}$, with $2\,\text{s}$ as the primary tolerance.

3.3.3 SETTING 3: COMMENTARY-TO-SPATIOTEMPORAL GROUNDING EVALUATION.

Setting 3 evaluates whether VLLMs can turn specified visual evidence into accurate and contextually appropriate sports commentary. The model receives chronologically ordered video frames, target bounding boxes, and preceding commentary, but neither box descriptions nor human reference commentary. Unlike Setting 2, it does not decide whether or when to comment; it must identify the relevant entities, actions, and competitive significance and describe them. We evaluate the generated commentary with GPT-6 Sol, reusing the commentary-quality criteria from Setting 2. The judge receives the human reference, target boxes and their visual-evidence annotations, and relevant video context. It assesses core events, entities, actions, outcomes, contextual consistency, relevance, informativeness, factual grounding, non-redundancy, clarity, and visual-evidence coverage.

4 EXPERIMENTS

4.1 SETTINGS

We first introduce the settings of the experiments briefly. More detailed settings are provided in the supplementary material.

Models. We evaluate two mainstream proprietary models on CG-SC-Bench: GPT-5.6-Sol (OpenAI, 2026) and Gemini-3.8-Flash (Google DeepMind, 2026). We also assess ten representative open-source multimodal models across different parameter scales.

Frame Sampling. We sample video frames at 1 FPS to preserve the temporal progression of each video. Specifically, frames at integer-second timestamps are extracted whenever available. If no frame exists at an exact integer timestamp, we select the nearest frame and record its actual timestamp. The same sampling strategy is applied consistently to all evaluated models.

Prompt. We design task-specific prompts for the three tasks. For **Visual Clue Localization**, we append the timestamp to each sampled frame and require the model to locate the visual evidence corresponding to the given commentary by returning its timestamp and bounding box. For **Streaming Commentary Generation**, the model is prompted to determine whether the current video content requires commentary and, if so, when the commentary should be generated and what it should describe. The model processes the video causally and is prohibited from using future frames. For **Clue-Conditioned Commentary Generation**, we provide the model with explicit visual clues in the form of timestamped bounding boxes and require it to generate commentary grounded in the specified objects and actions. All predictions are returned in predefined JSON formats to facilitate automatic evaluation.

4.2 MAIN RESULTS

As shown in Table 2, Gemini-3.8-Flash achieves the strongest overall performance. In Setting 1, its Acc@1s and P-STIoU reach 40.57% and 16.98%, compared with 20.97% and 7.36% for the best open-source model, Qwen3.8-27B. The much lower P-STIoU scores across models indicate that finding the approximate moment does not ensure accurate localization of the visual entity. In Setting 2, Gemini-3.8-Flash leads in trigger F1 (33.48%) and content score (40.73%), whereas Qwen3.5-35B-A3B leads in causal consistency (55.41%). Nemotron-3-Nano-30B-A3B has the lowest time error (0.53 s) but a trigger F1 of only 0.08%, showing why timing error must be interpreted alongside triggering performance.

In Setting 3, Gemini-3.8-Flash leads all four reported metrics, including weighted score (51.86%) and factual accuracy (21.30%). Qwen3.8-27B has the highest open-source weighted score (31.26%). Nevertheless, even the leading P-STIoU and factual accuracy scores remain low, leaving substantial room to improve visual grounding and factual reliability.

4.3 ANALYSIS

Fine-grained Spatiotemporal Grounding Analysis. Figure 7 shows each model's grounding performance at the video level. Each point represents one video, and the horizontal axis shows temporal accuracy (Acc@1s), while the vertical axis shows bounding-box overlap **among temporally**

Table 2: Performance of various open-source and closed-source MLLMs on CG-SC-Bench: accuracy within one second (Acc@1s), point-based spatiotemporal intersection over union (P-STIoU), trigger F1 (Trig. F1), timing error (T-Err.), content score (Cont.), causal consistency (Caus.), weighted score (W-Score), visual evidence (Vis. Evid.), core event (Core Ev.), and factual accuracy (Fact. Acc.). All metrics except T-Err. are reported as percentages; T-Err. is measured in seconds.

Models	Setting 1		Setting 2				Setting 3			
	Acc@1s	P-STIoU	Trig. F1	T-Err. (s)	Cont.	Caus.	W-Score	Vis. Evid.	Core Ev.	Fact. Acc.
Open-source MLLMs										
Qwen3.8-27B	20.97	7.36	31.43	0.90	33.02	49.17	31.26	20.34	30.74	6.27
Qwen3.5-9B	15.53	2.13	29.58	0.88	28.04	53.52	26.73	15.90	25.12	3.23
Qwen3.5-35B-A3B	18.11	4.00	30.50	0.95	35.29	**55.41**	27.98	16.77	26.16	3.62
Qwen3.5-27B	16.12	5.44	31.45	0.95	33.81	54.76	29.94	18.29	28.86	5.31
Nemotron-3-Nano-30B-A3B	5.71	0.87	0.08	**0.53**	11.32	–	20.76	8.88	15.31	4.66
InternVL3.5-30B-A3B	4.49	0.21	30.34	0.92	30.04	50.00	24.71	12.68	21.20	3.49
InternVL3.5-14B	1.71	0.16	27.31	0.81	26.82	43.51	19.50	8.39	12.69	2.09
GLM-4.6V-Flash	8.45	1.57	28.30	0.92	26.85	37.04	21.81	9.24	14.86	3.57
Gemma4-12B	13.95	1.29	31.26	0.92	26.41	51.35	20.12	7.64	12.64	4.44
Gemma4-26B-A4B	13.99	1.62	30.05	0.92	32.34	50.48	20.78	7.51	12.83	6.65
Closed-source MLLMs										
GPT5.6-sol	12.88	4.97	30.25	0.93	35.48	54.78	36.00	24.08	34.28	9.38
gemini-3.8-flash	**40.57**	**16.98**	**33.48**	0.96	**40.73**	52.26	**51.86**	**42.71**	**55.04**	**21.30**

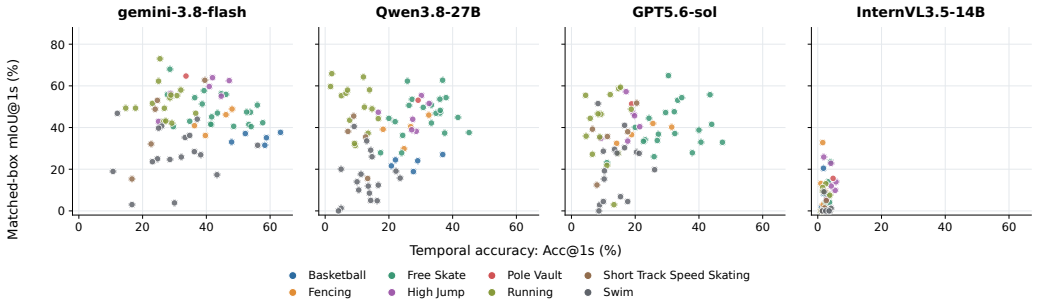


Figure 7: Video-level temporal and spatial grounding performance across 12 models.

matched predictions. Many models have points concentrated in the lower-left region, indicating difficulty with both temporal and spatial grounding. Some points instead show high overlap but low temporal accuracy, or the reverse. This variation shows that Some models locate the correct moment but draw inaccurate boxes, while others draw accurate boxes when the moment is correct but often select the wrong moment.

Gemini's video-level points generally lie farther to the right than other models, indicating stronger temporal grounding. Its Acc @1s reaches 40.57%, compared with 20.97% for the strongest open-source model, Qwen3.8-27B. Among temporally matched predictions, Gemini also achieves higher bounding-box overlap. Together, these gains give Gemini a Point-STIoU of 16.98%. Gemini's larger advantage lies in identifying the video moment described by the commentary. Its higher spatial overlap further improves the joint result. Gemini scores higher on both measures across all eight sports, suggesting that it more consistently aligns mentioned athletes and actions with the corresponding video moments across different events.

Commentary Triggering and Content Quality. Figure 8 presents the triggering and content results for Setting 2. In the left panel, InternVL3.5-14B achieves the highest recall (71.35%) but low precision (16.89%), resulting in a Trigger F1 of 27.31%. Its high recall comes with frequent unnecessary triggers. Gemini-3.8-Flash achieves the highest precision (23.83%) and Trigger F1 (33.48%), with a recall of 56.30%. Nemotron-3-Nano-30B-A3B illustrates the opposite failure mode: its recall of 0.04% and Trigger F1 of 0.08% show that it rarely triggers commentary when needed.

The right panel shows that triggering performance and content quality do not always improve together. Gemini-3.8-Flash leads on both Trigger F1 (33.48%) and Content score (40.73%). Gemma4-

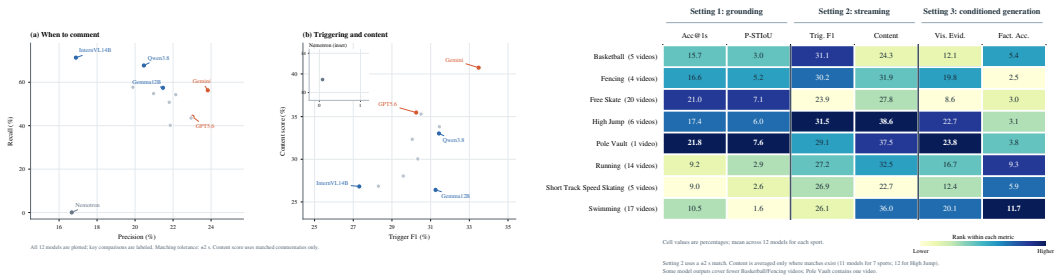


Figure 8: Commentary triggering and content quality (left) and performance across sports and settings (right).

12B scores slightly higher than GPT5.6-sol on Trigger F1 (31.26% versus 30.25%), while GPT5.6-sol scores substantially higher on content (35.48% versus 26.41%). This comparison shows that identifying appropriate moments and generating high-quality commentary are distinct challenges. Since Content score is calculated only for matched commentaries, it should be interpreted alongside Trigger F1.

As shown in Figure 8, the challenges vary by sport across the three settings. Free skating's single primary athlete makes grounding relatively easier, but continuous movements complicate commentary timing and technical descriptions. Splashes and underwater occlusion hinder athlete localization in swimming, while race progression may provide clearer commentary cues. High jump and pole vault have distinct event moments that aid triggering, yet details such as heights and attempt counts remain difficult to state accurately. These differences show that sports place different demands on visual grounding, commentary decisions, and factual generation.

5 Conclusion and Future Work

In this paper, we introduce CG-SC-Bench, a visual-clue-grounded benchmark for sports video commentary. Its three tasks evaluate whether models can ground commentary in visual evidence, decide when and what to say in a video stream, and generate commentary from specified visual clues. The dataset spans 8 major sports categories, more than 12 hours of video, and 2,684 commentary segments, including 2,168 with detailed visual clue annotations. Our evaluation reveals substantial room for improvement: even the strongest overall model, Gemini-3.8-Flash, achieves only 16.98% P-STIoU in grounding, 33.48% Trigger F1 in streaming commentary, and 21.30% factual accuracy in clue-conditioned generation. Future work will broaden the sports and commentary styles covered, strengthen evaluation of replays, multi-athlete interactions, and long-term context, and bring streaming evaluation closer to live broadcasts by requiring models to use only current and past information.

AI USE STATEMENT

Generative AI tools assisted with manuscript language editing and the development of data-processing and evaluation code. In Settings 2 and 3, we used LLM judges to assess generated commentary against predefined criteria. Human annotators created and reviewed the reference commentary and visual-evidence annotations. The authors reviewed the evaluation procedures, results, and manuscript claims and take full responsibility for the work.

REFERENCES

Satanjeev Banerjee and Alon Lavie. 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*, pp. 65–72, 2005.

Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 9887–9897. IEEE, 2023.

Kuangzhi Ge, Lingjun Chen, Kevin Zhang, Yulin Luo, Tianyu Shi, Liaoyuan Fan, Xiang Li, Guanqun Wang, and Shanghang Zhang. Scbench: A sports commentary benchmark for video llms. *arXiv preprint arXiv:2412.17637*, 2024.

Google DeepMind. Gemini 3.8 Flash: Model card. `https://deepmind.google/models/model-cards/gemini-3-8-flash/`, 2026. Accessed: 2026-09-24.

Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 13516–13525. IEEE, 2021.

Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pp. 74–81, 2004.

Zhaoyu Liu, Xi Weng, Lianyu Hu, Zhe Hou, Kan Jiang, Jin Song Dong, and Yang Liu. Tennisexpert: towards expert-level analytical sports video understanding. *arXiv preprint arXiv:2603.13397*, 2026.

Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-caption: Dense video captioning for soccer broadcasts commentaries. In *2023 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 5074–5085. IEEE, 2023.

OpenAI. GPT-5.6: Frontier intelligence that scales with your ambition. `https://openai.com/index/gpt-5-6/`, 2026. Accessed: 2026-09-24.

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

Ji Qi, Jifan Yu, Teng Tu, Kunyu Gao, Yifan Xu, Xinyu Guan, Xiaozhi Wang, Bin Xu, Lei Hou, Juanzi Li, et al. Goal: A challenging knowledge-grounded video captioning benchmark for real-time soccer commentary generation. In *Proceedings of the 32nd ACM international conference on information and knowledge management*, pp. 5391–5395, 2023.

Jiayuan Rao, Haoning Wu, Chang Liu, Yanfeng Wang, and Weidi Xie. Matchtime: Towards automatic soccer game commentary generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 1671–1685, 2024.

Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2613–2622. IEEE, 2020.

Zongheng Tang, Yue Liao, Si Liu, Guanbin Li, Xiaojie Jin, Hongxu Jiang, Qian Yu, and Dong Xu. Human-centric spatio-temporal video grounding with visual transformers. *IEEE Transactions on Circuits and Systems for Video Technology*, 32(12):8238–8249, 2021.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575, 2015.

Kaiwen Wang, Kaili Zheng, Rongrong Deng, Yiming Shi, Chenyi Guo, and Ji Wu. Boxcomm: Benchmarking category-aware commentary generation and narration rhythm in boxing. *arXiv preprint arXiv:2604.04419*, 2026.

Dekun Wu, He Zhao, Xingce Bao, and Richard P Wildes. Sports video analysis on large-scale data. In *European conference on computer vision*, pp. 19–36. Springer, 2022.

Lifang Wu, Yuyang Wu, Yangdong Gao, Fengyu Liu, Ya Jing, and Liang Wang. Nba_streaming: A large-scale benchmark for fine-grained basketball commentary generation in continuous streams. *arXiv preprint arXiv:2608.09200*, 2026.

Tao Wu, Runyu He, Gangshan Wu, and Limin Wang. Sportshhi: A dataset for human-human interaction detection in sports videos. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 18537–18546. IEEE, 2024.

Zeyu Xi, Haoying Sun, Yaofei Wu, Junchi Yan, Haoran Zhang, Lifang Wu, Liang Wang, and Changwen Chen. Player-centric multimodal prompt generation for large language model based identity-aware basketball video captioning. In *2025 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 24330–24339. IEEE, 2025.

Haotian Xia, Zhengbang Yang, Junbo Zou, Rhys Tracy, Yuqing Wang, Chi Lu, Christopher Lai, Yanjun He, Xun Shao, Zhuoqing Xie, et al. Sportu: A comprehensive sports understanding benchmark for multimodal large language models. In *International Conference on Learning Representations*, volume 2025, pp. 29892–29950, 2025.

Haotian Xia, Haonan Ge, Junbo Zou, Hyun Choi, Xuebin Zhang, Danny Suradja, Botao Rui, Ethan Tran, Wendy Jin, Zhen Ye, et al. Sportr: A benchmark for multimodal large language model reasoning in sports. In *International Conference on Learning Representations*, volume 2026, pp. 22777–22804, 2026.

Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 2939–2948. IEEE, 2022.

Yichen Xu, Yuanhang Liu, Chuhan Wang, Zihan Zhao, Jianzhe Ma, Wenxuan Wang, Qin Jin, et al. Refereebench: Are video mllms ready to be multi-sport referees. *arXiv preprint arXiv:2604.15736*, 2026.

Jiali Yao, Xin Gu, Xinran Deng, Mengrui Dai, Bing Fan, Zhipeng Zhang, Yan Huang, Heng Fan, and Libo Zhang. Omnistvg: Toward spatio-temporal omni-object video grounding. In *International Conference on Learning Representations*, volume 2026, pp. 14584–14605, 2026.

Benhui Zhang, Junyu Gao, and Yuan Yuan. A descriptive basketball highlight dataset for automatic commentary generation. In *Proceedings of the 32nd ACM international conference on multimedia*, pp. 10316–10325, 2024.

Zhu Zhang, Zhou Zhao, Yang Zhao, Qi Wang, Huasheng Liu, and Lianli Gao. Where does it exist: Spatio-temporal video grounding for multi-form sentences. In *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 10665–10674. IEEE, 2020.

Junbo Zou, Haotian Xia, Zhen Ye, Shengjie Zhang, Christopher Lai, Vicente Ordonez, Weining Shen, and Hanjie Chen. Deepsport: A multimodal large language model for comprehensive sports video reasoning via agentic reinforcement learning. *arXiv preprint arXiv:2511.12908*, 2025.

A Dataset Details

A.1 Dataset Composition

CG-SC-Bench covers 8 major sports categories and 27 subcategories, comprising 72 videos and 2,684 commentary segments with a total duration of 12.44 hours. As shown in Table 3, the dataset contains an average of 3.60 commentary segments per minute. This frequency provides repeated opportunities to evaluate whether models can decide when an event warrants commentary and what to say.

Table 3: Annotation statistics, detailing the number of commentary segments, grounded segments and boxes.

Annotation Statistics	
#Total Grounded Commentary Segments	2168
#Total Annotation Boxes	4943
#Avg. commentary segments per video	37.28
#Avg. grounded segments per video	30.11
#Avg. boxes per grounded segment	2.28
#Avg. commentary segments per minute	3.60
#Avg. grounded segments per minute	2.91
#Avg. boxes per minute	6.64

Beyond commentary timing, the dataset provides explicit visual grounding for 2,168 segments (80.8%), which are associated with 4,943 bounding boxes. This connection enables generated commentary to be assessed against localized visual evidence as well as the reference text. With an average of 2.28 boxes per grounded segment, the annotations also support evaluation of whether a model accounts for the visual evidence associated with each segment.

The remaining 516 segments are retained in chronological order, although they lack detailed bounding-box annotations. Keeping them alongside the grounded segments preserves the progression of the original commentary, allowing model responses to be evaluated within the ongoing competition context.

A.2 Video Collection and Selection

Videos were primarily collected from YouTube. We selected competition broadcasts with clear, intelligible commentary, synchronized audio and video, and relatively frequent commentary focused on the ongoing competition. This provided sufficient material to annotate how events were described over time.

During screening, we excluded videos with frequent interruptions or stuttering in the commentary, noticeable audio-video desynchronization, or commentary that was largely unrelated to the competition shown. These criteria helped ensure reliable temporal alignment between commentary and footage for the benchmark tasks.

As shown in Figures 9–12, video length and commentary timing vary considerably across CG-SC-Bench. The 72 videos range from 4.37 to 31.24 minutes in length, with commentary densities ranging from 2.08 to 5.97 segments per minute. Even videos of similar length can differ in commentary density, so video duration alone does not determine how often commentary occurs. At a finer temporal scale, the median gap between adjacent commentary segments is 1.48 seconds, yet 278 of the 2,612 gaps are at least 10 seconds long. Extended intervals without annotated commentary occur alongside closely spaced segments. **Predicting when to generate commentary is therefore essential to streaming sports commentary**: a model must decide when to respond as the competition unfolds, rather than produce commentary at a fixed rate.

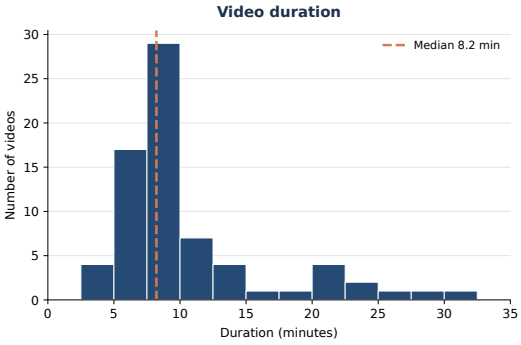


Figure 9: Distribution of video durations.

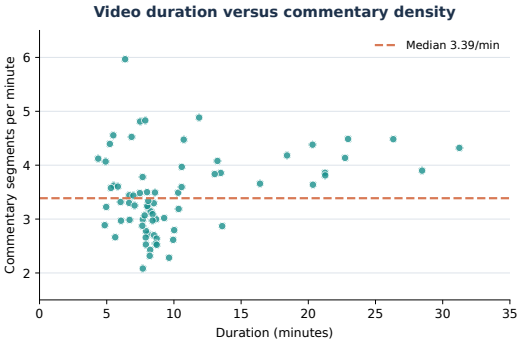


Figure 10: Video duration and commentary density.

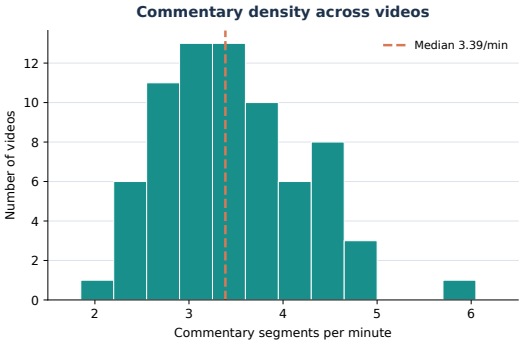


Figure 11: Distribution of commentary density.

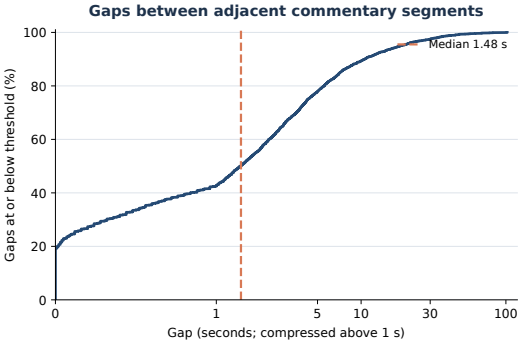


Figure 12: Gaps between adjacent commentary segments.

Table 4: Fields in the commentary annotation format.

Field	Description
`start`, `end`	Start and end times of the commentary utterance in the original video.
`text`	Full transcript of the commentary utterance.
`bbox`	List of visual evidence boxes associated with the utterance. The list may be empty or absent when no identifiable visual evidence is available.
`frame`	Video frame number on which the box is annotated.
`time`	Actual timestamp of the annotated frame in the original video, in seconds.
`x`, `y`	Position of the box's upper-left corner, expressed as percentages of the frame width and height.
`width`, `height`	Size of the box, expressed as percentages of the frame width and height.
`rotation`	Rotation angle of the box; zero indicates no rotation.
`description`	Commentary text associated with the box. Angle brackets identify the entity enclosed by the box.

A.3 Annotation Guidelines and Examples

Annotation Format. Each annotation is organized around a commentary utterance and records its start and end times in the video, together with the transcribed text. When part of an utterance corresponds to visible evidence in the video, we associate it with one or more bounding boxes. Each box records its video frame and timestamp, its position and size as percentages of the frame dimensions, and the corresponding phrase from the commentary. An utterance may have multiple boxes to cover different participants, objects, or moments mentioned in the commentary. For utterances without identifiable visual evidence, we retain the text and timestamps without associating any boxes.

Listing 1: An example of a commentary annotation.


```
{
  "start": "00:11:46,380",
  "end": "00:11:53,620",
  "text": "Durant finds Roberson. and it's passed to Adams sails Out of
      bounds. Iguodala is going to come back in",
  "bbox": [
    {
      "frame": 17657,
      "rotation": 0,
      "x": 61.56250000000001,
      "y": 34.22222222222225,
      "width": 5.499999999999971,
      "height": 26.44444444444446,
      "time": 706.28,
      "description": "<Durant> finds Roberson. and it's passed to Adams
          sails Out of bounds."
    },
    {
      "frame": 17668,
      "rotation": 0,
      "x": 27.687499999999993,
      "y": 34.66666666666659,
      "width": 6.749999999999999,
      "height": 26.6666666666667,
      "time": 706.72,
      "description": "Durant finds <Roberson>. and it's passed to Adams
          sails Out of bounds."
    },
    {
      "frame": 17688,
      "rotation": 0,
      "x": 24.9375,
      "y": 38.22222222222222,
      "width": 8.875,
      "height": 26.88888888888889,
      "time": 707.52,
      "description": "Durant finds Roberson. and it's passed to <Adams>
          sails Out of bounds."
    }
  ]
}
```


Listing 1 shows the format of a commentary annotation, and Table 4 defines its fields. Each annotation records the commentary text and its start and end times. When part of the commentary corresponds to visible evidence, the ‘bbox‘ field links it to one or more boxes.

Annotation Examples The following examples illustrate how we annotate different relationships between commentary and visual evidence.

(1) Athlete introduction. When commentary introduces a visible athlete, we annotate the athlete and link the box to the corresponding mention in the commentary, as shown in Fig. 13 and Fig. 14.

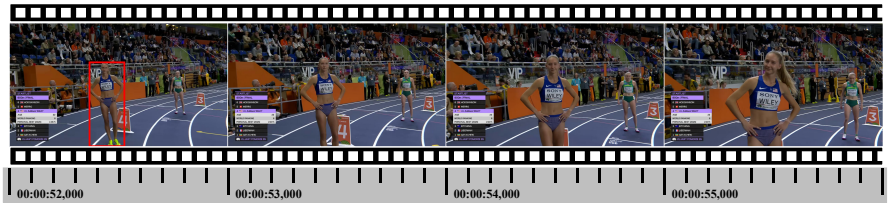


Figure 13: Annotation example 1 (Athlete introduction).

Listing 2: JSON annotation for the example in Fig. 13.


```
{
  "start": "00:00:52,200",
  "end": "00:01:05,680",
  "text": "Addison Wiley, North American record holder over 1,000 m,
      fifth in the world juniors a few years ago, but this is the
      biggest race of her life.",
  "bbox": [
    {
      "frame": 1323,
      "rotation": 0,
      "x": 36.203319502074685,
      "y": 31.005071461502997,
      "width": 16.59751037344398,
      "height": 68.97187644075612,
      "time": 52.92,
      "description": "Addison Wiley, North American record holder over
          1,000 m, fifth in the world juniors a few years ago, but this
          is the biggest race of her life."
    }
  ]
}
```

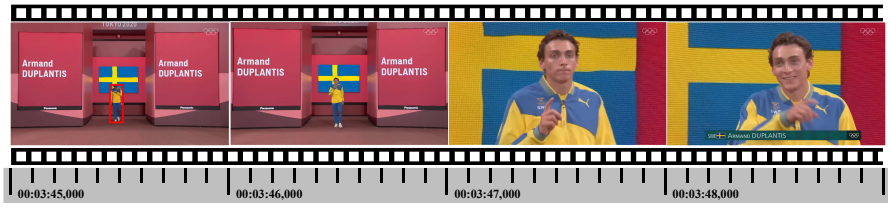



Figure 14: Annotation example 2 (Athlete introduction).

Listing 3: JSON annotation for the example in Fig. 14.


```
{
  "start": "00:03:45,430",
  "end": "00:03:53,300",
  "text": "and Here he is the great Armand Duplantis, just 21 years of
      age, rank No.1 in the world, season best of 6.10m.",
  "bbox": [
    {
      "frame": 5639,
      "rotation": 0,
      "x": 45.22821576763486,
      "y": 49.99999999999987,
      "width": 6.742738589211552,
      "height": 31.535269709543517,
      "time": 225.56,
      "description": "and Here he is the great Armand Duplantis, just 21
          years of age, rank No.1 in the world, season best of 6.10m."
    }
  ]
}
```


(2) Continuous Actions. A single utterance may describe several actions performed in sequence. We annotate the athlete at the moments corresponding to those actions, preserving their temporal order rather than using one frame to represent the entire sequence, as shown in Fig. 15.

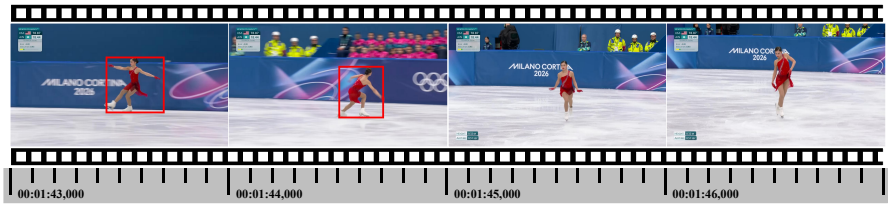


Figure 15: Annotation example 3 (Continuous Actions).

Listing 4: JSON annotation for the example in Fig. 15.


```
{
  "start": "00:01:43,470",
  "end": "00:01:45,350",
  "text": "Good footwork to Triple Lutz.",
  "bbox": [
    {
      "frame": 2512,
      "rotation": 0,
      "x": 43.648208469055376,
      "y": 27.029798528169863,
      "width": 26.92725298588491,
      "height": 45.74737604053565,
      "time": 100.48,
      "description": "<Good footwork> to Triple Lutz."
    },
    {
      "frame": 2547,
      "rotation": 0,
      "x": 50.16286644951141,
      "y": 34.557847750030184,
      "width": 20.73832790445162,
      "height": 42.27289178429244,
      "time": 101.88,
      "description": "Good footwork to <Triple Lutz>."
    }
  ]
}
```


(3) Athlete Comparison. In events such as running, commentary often compares athletes' positions or describes one athlete gaining on another. We annotate the athletes separately at relevant moments so that their relative positions and changes over time can be identified, as shown in Fig. 16.

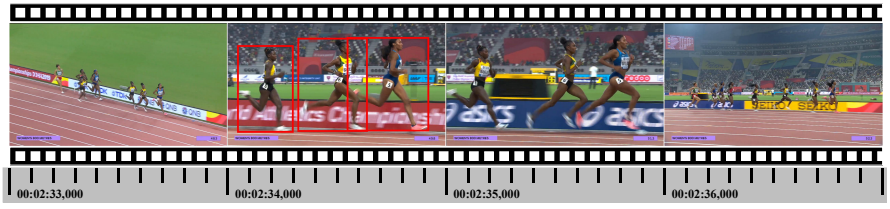


Figure 16: Annotation example 4 (Athlete Comparison).

Listing 5: JSON annotation for the example in Fig. 16.


```
{
  "start": "00:02:26,620",
  "end": "00:02:42,400",
  "text": "Jenny Meadows. that was quick through the first 200 meters so
      27 seconds is incredibly fast but she's followed by Goule and
     Nakaayi they all like to lead these women.",
```

```
  "bbox": [
    {
      "frame": 3864,
      "rotation": 0,
      "x": 54.61454940282303,
      "y": 11.394619375075417,
      "width": 37.67643865363735,
      "height": 76.63168054047529,
      "time": 154.56,
      "description": "but <she>'s followed by Goule and Nakaayi they all
          like to lead these women."
    },
    {
      "frame": 3864,
      "rotation": 0,
      "x": 32.13897937024973,
      "y": 11.973700084449272,
      "width": 32.13897937024973,
      "height": 76.43865363735071,
      "time": 154.56,
      "description": "but she's followed by <Goule> and Nakaayi they all
          like to lead these women."
    },
    {
      "frame": 3864,
      "rotation": 0,
      "x": 4.560260586319218,
      "y": 17.957534081312602,
      "width": 25.841476655808908,
      "height": 71.2269272529859,
      "time": 154.56,
      "description": "but she's followed by Goule and <Nakaayi> they all
          like to lead these women."
    }
  ]
}
```


(4) Athlete Interaction. When commentary describes an interaction between athletes, we annotate the participants at the relevant moments and link each box to its corresponding mention in the commentary. This allows the visual evidence to capture their roles and how the interaction unfolds over time, as shown in Fig. 17.

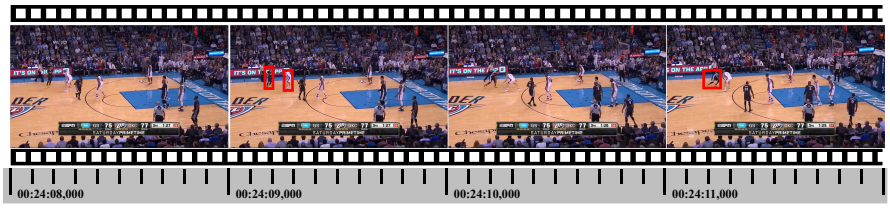


Figure 17: Annotation example 5 (Athlete Interaction).

Listing 6: JSON annotation for the example in Fig. 17.


```
{
  "start": "00:24:07,920",
  "end": "00:24:14,270",
  "text": "Singler now on Curry. crossover. step back. 3 pointer. bang",
  "bbox": [
    {
      "frame": 36229,
      "rotation": 0,
      "x": 24.0625,
```

```
      "y": 35.77777777777777,
      "width": 5,
      "height": 18.444444444444443,
      "time": 1449.16,
      "description": "<Singler> now on Curry."
    },
    {
      "frame": 36229,
      "rotation": 0,
      "x": 15.312500000000007,
      "y": 33.11111111111107,
      "width": 4.874999999999996,
      "height": 19.555555555555543,
      "time": 1449.16,
      "description": "Singler now on <Curry>."
    },
    {
      "frame": 36268,
      "rotation": 0,
      "x": 16.562500000000007,
      "y": 35.55555555555561,
      "width": 8.875000000000014,
      "height": 16.666666666666696,
      "time": 1450.72,
      "description": "crossover. step back. 3 pointer. bang"
    }
  ]
}
```


(5) Multiple-Athlete Introduction. When an utterance introduces several athletes together, we annotate each identifiable athlete separately. This allows one utterance to be associated with multiple visual objects while retaining the correspondence between each athlete and the commentary text, as shown in Fig. 18.

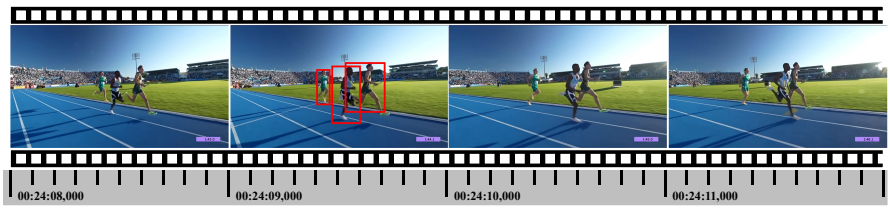


Figure 18: Annotation example 6 (Multiple-Athlete Introduction).

Listing 7: JSON annotation for the example in Fig. 18.


```
{
  "start": "00:05:37,010",
  "end": "00:05:46,589",
  "text": "And Australia not out of the running yet. Thomas Reynolds has
      the baton for them. The medals are surely between these three.
     They're miles clear of the rest at the moment",
  "bbox": [
    {
      "frame": 8575,
      "rotation": 0,
      "x": 39.31250000000001,
      "y": 35.833333333333336,
      "width": 6.5625,
      "height": 29.33333333333333,
      "time": 343,
      "description": "The medals are surely between these three. They're
         miles clear of the rest at the moment"
```

```
    },
    {
      "frame": 8575,
      "rotation": 0,
      "x": 46.25000000000002,
      "y": 32.99999999999999,
      "width": 13.499999999999968,
      "height": 47.333333333333385,
      "time": 343,
      "description": "The medals are surely between these three. They're
          miles clear of the rest at the moment"
    },
    {
      "frame": 8575,
      "rotation": 0,
      "x": 52.343750000000135,
      "y": 30.166666666666654,
      "width": 18.656249999999876,
      "height": 41.16666666666667,
      "time": 343,
      "description": "The medals are surely between these three. They're
          miles clear of the rest at the moment"
    }
  ]
}
```


(6) Retrospective Commentary. Commentary may refer back to an earlier action while the current frame shows something else. If the referenced action can be located in the video, we associate the commentary with visual evidence from that earlier moment. Otherwise, we retain the commentary text and timestamps without assigning a bounding box, as shown in Fig. 19.

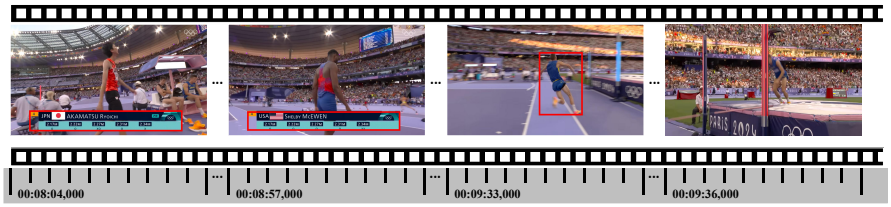


Figure 19: Annotation example 7 (Retrospective Commentary).

Listing 8: JSON annotation for the example in Fig. 19.


```
{
   "start": "00:09:35,960",
   "end": "00:09:55,700",
   "text": "No, it really looked like it. So now McEwen, Akamatsu and
       Doroshchuk have all gone at the 2.34 and have all missed twice.
       Barshim, Sottile and Kerr there, if they all miss again we know
       our medalists. It's just a matter of which order. such is this
       sport.",
   "bbox": [
     {
       "frame": 14868,
       "rotation": 0,
       "x": 9.691160809371672,
       "y": 77.07371908649864,
       "width": 77.63578274760383,
       "height": 18.364690569163418,
       "time": 594.72,
       "description": "So now McEwen, Akamatsu and <Doroshchuk> have all
           gone at the 2.34 and have all missed twice."
     },
```

```
    {
      "frame": 13448,
      "rotation": 0,
      "x": 9.691160809371672,
      "y": 77.831025914093,
      "width": 77.74227902023428,
      "height": 17.418057034670454,
      "time": 537.92,
      "description": "So now <McEwen>, Akamatsu and Doroshchuk have all
          gone at the 2.34 and have all missed twice."
    },
    {
      "frame": 12121,
      "rotation": 0,
      "x": 9.691160809371672,
      "y": 77.07371908649864,
      "width": 77.84877529286474,
      "height": 18.743343982960603,
      "time": 484.84,
      "description": "So now McEwen, <Akamatsu> and Doroshchuk have all
          gone at the 2.34 and have all missed twice."
    }
  ]
}
```


B More Details in Evaluation

B.1 Setting 1

In Setting 1, we evaluate whether MLLMs can localize the visual clues that support a given sports commentary within the corresponding video segment, thereby assessing fine-grained understanding of athletes, actions, and their spatiotemporal relationships.

For each manually annotated commentary, we determine the corresponding video interval and sample the video at 1 FPS. Given the fine-grained commentary and the ordered video frames, the model is required to identify the relevant visual evidence, determine when it appears, and localize the associated visual entities. Each prediction therefore contains both a timestamp and one or more bounding boxes. When a commentary involves multiple athletes, actions, or object relations, multiple visual clues may be returned.

The prompt instructs the model to localize only the visual evidence that directly supports the commentary, preserve the input text, and return predictions in a predefined structured format for automatic parsing and evaluation.

For evaluation, we consider both temporal and spatial localization. Temporal performance is measured by comparing predicted clue timestamps with manual annotations, while spatial performance is measured by the overlap between predicted and ground-truth bounding boxes. When multiple visual entities are involved, predictions and annotations are matched according to their spatiotemporal correspondence before computing the localization metrics.

In Setting 1, we further provide a more comprehensive view of the experimental results. Specifically, we report the complete performance of all evaluated models across the selected localization metrics, enabling a more detailed comparison of their temporal and spatial grounding capabilities

As shown in Table 5, the results of Setting 1 reveal clear differences among models in visual clue grounding. Gemini-3.8-Flash achieves the best overall performance across nearly all metrics, with higher Acc@0.5s, Acc@1.0s, and Acc@2.0s scores as well as the lowest MTE, indicating more accurate temporal localization of relevant visual evidence. It also leads in mIoU@1.0s and STIoU under different temporal tolerances, demonstrating stronger spatial localization and overall spatiotemporal grounding capability. In comparison, some models show competitive spatial localization performance, for example, GPT-5.6-Sol and Qwen3.8-27B. They both obtain relatively high mIoU scores, but still lag behind in temporal localization accuracy. Overall, these results suggest that es-

Table 5: Performance comparison on Setting 1: Visual Clue Grounding.

Model	Acc@0.5s	Acc@1.0s	Acc@2.0s	MTE	mIoU@1.0s	STIoU@0.5s	STIoU@1.0s	STIoU@2.0s
Qwen3.8-27B	0.114	0.210	0.326	8.44	0.351	0.042	0.074	0.108
Qwen3.5-9B	0.071	0.155	0.271	13.24	0.137	0.009	0.021	0.038
Qwen3.5-35B-A3B	0.093	0.181	0.294	7.77	0.221	0.021	0.040	0.063
Qwen3.5-27B	0.086	0.161	0.284	10.08	0.338	0.028	0.054	0.092
Nemotron-3-Nano-30B-A3B	0.024	0.057	0.112	11.70	0.152	0.003	0.009	0.016
InternVL3.5-30B-A3B	0.024	0.045	0.079	15.70	0.046	0.001	0.002	0.003
InternVL3.5-14B	0.008	0.017	0.031	40.65	0.091	0.001	0.002	0.003
GLM-4.6V-Flash	0.037	0.085	0.155	12.82	0.185	0.006	0.016	0.028
GPT-5.6-Sol	0.072	0.129	0.201	9.53	0.386	0.027	0.050	0.072
Gemma4-12B	0.065	0.140	0.243	11.93	0.092	0.006	0.013	0.022
Gemma4-26B-A4B	0.073	0.140	0.232	12.41	0.116	0.008	0.016	0.027
Gemini-3.8-Flash	**0.238**	**0.406**	**0.541**	**3.17**	**0.419**	**0.101**	**0.170**	**0.211**

tablishing precise spatiotemporal correspondences between sports commentary and concrete visual evidence remains challenging, especially when accurate temporal and spatial grounding must be achieved simultaneously.

B.2 SETTING 2

In Setting 2, we evaluate the causal streaming commentary capability of MLLMs on continuous sports videos. Unlike offline understanding based on complete videos, this setting requires the model to make decisions using only the current and previously observed video content, without access to future frames. As the video stream progresses, the model must autonomously determine whether commentary is needed, when it should be generated, and what content should be produced, thereby evaluating its online understanding, triggering decisions, and real-time commentary capability.

For each video, sampled frames are continuously provided to the model in chronological order, together with the model's previously generated commentaries as historical context. All videos are uniformly sampled at 1 FPS and processed through consecutive temporal windows. At any time, the model can access only the current window and previously observed visual content, ensuring causal inference throughout the evaluation.

For each input, the model first determines whether the current video content warrants commentary. Further, the model need to determines an appropriate commentary time and generates the corresponding text. Previously generated commentaries are included in subsequent inputs, allowing the model to maintain contextual understanding of the ongoing game while reducing repetitive outputs.

For evaluation, we consider three aspects: commentary triggering, content quality, and causal consistency. Precision, Recall, Trigger F1, and Time Error are used to evaluate whether commentary is generated at appropriate times. Precision and Recall measure the accuracy and coverage of commentary triggers, respectively, while Trigger F1 jointly considers both. Time Error measures the average temporal deviation between successfully matched predictions and human annotations. For matched commentaries, Content Score and Rubric F1 are further used to assess generation quality, while Causal Consistency and Causal Rubric F1 evaluate whether the generated commentary is consistent with the visual information observed so far and avoids relying on future content. The primary temporal matching tolerance in Setting 2 is set to ±2 seconds.

In addition, we use GPT-6-Sol as a unified judge to evaluate the quality of generated commentary. Each predicted commentary is first matched with the corresponding ground-truth commentary according to the predefined temporal matching rule, and only successfully matched prediction-reference pairs are submitted for quality evaluation. For each matched pair, the judge receives both the generated commentary and the corresponding ground-truth commentary and evaluates the prediction according to ten predefined criteria: Core Event, Entities, Action, Outcome, Context, Relevance, Informativeness, Factuality, Non-redundancy, and Clarity. For each criterion, the judge outputs one of three labels: MET, UNMET, or NA, where MET indicates that the criterion is satisfied, UNMET indicates that it is not satisfied, and NA indicates that the criterion is not applicable to the current sample. The score for each criterion is computed as $\mathrm{MET}/(\mathrm{MET}+\mathrm{UNMET})$. These criterion-level judgments are further combined according to their predefined weights to compute the overall commentary quality score. The ten evaluation criteria, together with their definitions and weights, are summarized in Table 6.

Table 6: Evaluation criteria used for commentary quality assessment in Setting 2.

Metric	Weight	Description
Core Event	10	Captures the main event or point of the reference.
Entities	9	Correctly identifies key athletes, teams, or other entities.
Action	9	Accurately describes the principal sports action.
Outcome	8	Correctly describes the result, consequence, or game state.
Context	5	Consistent with the surrounding competition context.
Relevance	5	Focuses on the annotated event rather than unrelated content.
Informativeness	4	Provides specific and useful information.
Factuality	10	Avoids contradictory or unsupported factual claims.
Non-redundancy	4	Adds new information rather than repeating prior commentary.
Clarity	2	Clear, coherent, and suitable as sports commentary.

Table 7: Performance comparison on Setting 2: Causal Streaming Commentary Decision.

Model	Precision	Recall	Trigger F1	Time Error	Content Score	Rubric F1	Causal Cons.	Causal Rubric F1
Qwen3.8-27B	0.205	0.677	0.314	0.900	0.330	0.104	0.492	0.091
Qwen3.5-9B	0.199	0.577	0.296	0.883	0.280	0.083	0.535	0.075
Qwen3.5-35B-A3B	0.218	0.508	0.305	0.955	0.353	0.108	**0.554**	0.094
Qwen3.5-27B	0.221	0.543	0.315	0.946	0.338	0.106	0.548	0.094
Nemotron-3-Nano-30B-A3B	0.167	0.000	0.001	**0.530**	0.113	0.000	–	0.000
InternVL3.5-30B-A3B	0.210	0.548	0.303	0.924	0.300	0.091	0.500	0.080
InternVL3.5-14B	0.169	**0.713**	0.273	0.814	0.268	0.073	0.435	0.066
GLM-4.6V-Flash	0.218	0.402	0.283	0.916	0.269	0.076	0.370	0.069
GPT-5.6-Sol	0.230	0.441	0.303	0.926	0.355	0.107	0.548	0.095
Gemma4-12B	0.215	0.575	0.313	0.920	0.264	0.083	0.514	0.076
Gemma4-26B-A4B	0.229	0.435	0.300	0.919	0.323	0.097	0.505	0.088
Gemini-3.8-Flash	**0.238**	0.563	**0.335**	0.959	**0.407**	**0.136**	0.523	**0.116**

As shown in Table 7, the complete results reveal clear differences among models in deciding when to comment and in the quality of the generated commentary. Some models achieve relatively high Recall, indicating broader coverage of potentially comment-worthy events, but do not necessarily maintain high Precision. In contrast, other models produce higher-quality commentary while missing more events that should be commented on. Gemini-3.8-Flash performs strongly on Trigger F1, Content Score, Rubric F1, and Causal Rubric F1, indicating a favorable balance between trigger decisions and commentary quality. Meanwhile, Qwen3.5-35B-A3B achieves a relatively high Causal Consistency score, suggesting stronger consistency under the causal constraint.

It is worth noting that although Nemotron-3-Nano-30B-A3B obtains the lowest Time Error, this result should not be interpreted independently as evidence of superior temporal performance. The model generates only 6 predictions over the 71 evaluated videos, with a Recall of approximately 0.0004 and a Trigger F1 close to zero. Consequently, its Time Error is computed from only a very small number of successfully matched predictions. The low Time Error therefore reflects the limited set of matched samples rather than stable and accurate streaming trigger performance. Time Error should thus be interpreted jointly with Precision, Recall, and Trigger F1.

B.3 Setting 3

In Setting 3, we evaluate the ability of models to generate sports commentary under explicit visual conditions. By directly providing the model with target visual clues related to the current event, the model is required to generate corresponding sports commentary based on this visual evidence. This setting is designed to further separate visual localization ability from the ability to generate language grounded in visual evidence, allowing us to more directly assess whether models can understand explicitly provided fine-grained visual information and transform it into accurate, relevant, and context-appropriate sports commentary.

For each commentary sample with manually annotated visual information, we first determine its corresponding video interval and extract the target bounding boxes and their descriptions associated with the commentary. The model input consists of temporally ordered sampled frames, target visual clues, and necessary contextual information. Since the target visual evidence is explicitly provided,

Table 8: Performance of various open-source and closed-source MLLMs on Setting 3: weighted score (W-Score), core event (Core Ev.), entities (Ent.), action (Act.), outcome (Out.), context (Ctx.), relevance (Rel.), informativeness (Info.), factuality (Fact.), non-redundancy (Non-red.), clarity (Clar.), and visual evidence (Vis. Evid.). All metrics are reported as scores, with higher values indicating better performance.

Model	W-Score	Core Ev.	Ent.	Act.	Out.	Ctx.	Rel.	Info.	Fact.	Non-red.	Clar.	Vis. Evid.
Qwen3.8-27B	0.313	0.307	0.408	0.272	0.245	0.499	0.499	0.359	0.063	–	0.996	0.203
Qwen3.5-9B	0.267	0.251	0.333	0.233	0.193	0.475	0.462	0.305	0.032	1.000	0.984	0.159
Qwen3.5-35B-A3B	0.280	0.262	0.388	0.242	0.204	0.463	0.469	0.315	0.036	1.000	0.998	0.168
Qwen3.5-27B	0.299	0.289	0.395	0.257	0.223	0.520	0.483	0.342	0.053	–	0.996	0.183
Nemotron-3-Nano-30B-A3B	0.208	0.153	0.252	0.161	0.111	0.427	0.419	0.184	0.047	–	0.993	0.089
InternVL3.5-30B-A3B	0.247	0.212	0.298	0.224	0.173	0.460	0.500	0.224	0.035	–	0.999	0.127
InternVL3.5-14B	0.195	0.127	0.259	0.152	0.111	0.507	0.338	0.143	0.021	–	0.999	0.084
GLM-4.6V-Flash	0.218	0.149	0.271	0.166	0.133	0.612	0.384	0.148	0.036	–	0.987	0.092
GPT-5.6-Sol	0.360	0.343	0.464	0.327	0.261	0.622	0.623	0.366	0.094	1.000	**1.000**	0.241
Gemma4-12B	0.201	0.126	0.274	0.138	0.130	0.512	0.356	0.121	0.044	–	0.999	0.076
Gemma4-26B-A4B†	0.208	0.128	0.271	0.137	0.133	0.599	0.331	0.096	0.066	–	**1.000**	0.075
Gemini-3.8-Flash	**0.519**	**0.550**	**0.632**	**0.498**	**0.407**	**0.712**	**0.783**	**0.593**	**0.213**	–	0.996	**0.427**

this task reduces the influence of failures in visual clue retrieval and focuses more directly on the model's ability to understand fine-grained visual information and convert it into natural-language commentary. For commentary quality evaluation, Setting 3 follows the same evaluation criteria and scoring scheme as Setting 2. We use GPT-6-Sol as the unified judge and evaluate each generated commentary according to ten criteria.

In addition, we separately report the Visual Evidence metric to evaluate whether the generated commentary is genuinely supported by the provided visual clues. This metric differs from Relevance: Relevance mainly measures whether the generated content focuses on the target event, whereas Visual Evidence emphasizes whether the specific statements in the generated commentary can be directly supported by the provided target visual evidence. It therefore provides a more direct measure of whether the model actually uses the explicit visual conditions rather than generating plausible commentary based only on competition context or language priors.

The complete results for Setting 3 are shown in Table 8. Overall, the evaluated models exhibit clear differences across the fine-grained criteria. Some models are able to generate relatively fluent and well-structured sports commentary, yet still make substantial errors in identifying key entities, actions, event outcomes, and factual details. In particular, Clarity scores are generally high across models, whereas Factuality scores are considerably lower. This suggests that linguistic fluency does not necessarily indicate that a model has correctly understood and utilized the provided visual evidence. Even when target bounding boxes and visual clues are explicitly provided, models may still generate specific claims that are unsupported by the visual evidence or inconsistent with the actual competition event.

It should also be noted that the Non-redundancy criterion contains a large number of NA samples in Setting 3. Since most samples do not provide sufficient historical commentary context for evaluating repetition, this criterion can only be assessed on a small subset of applicable samples. Therefore, its results should be interpreted together with the number of valid samples and should not be used independently as a primary indicator for comparing overall model performance.

C VISUALIZATION OF MODEL OUTPUTS IN THE THREE SETTINGS

C.1 SETTING1

Input:



"text": "Durant back out, Westbrook. Westbrook attacks to Ibaka. 2 pointer, knocks it in.",
"description": "<Westbrook> attacks to Ibaka."

Ground-Truth:



```
"x": 45.79,
"y": 44.88,
"width": 7.02,
"height": 22.71,
"time": 1379.68,
"description": "<Westbrook> attacks to Ibaka."
```

```
"x": 44.72,
"y": 37.69,
"width": 7.13,
"height": 28.02,
"time": 1379.68,
"description": "Westbrook attacks to <Ibaka>."
```

Qwen3.8-27B



```
"bbox_2d": [638.0, 372.0, 698.0, 598.0],
time": 1377.0,
"description": "<Durant> back out, Westbrook."
```

```
"bbox_2d": [638.0, 375.0, 695.0, 595.0],
"time": 1377.0,
"description": "<Westbrook> attacks to Ibaka."
```

Figure 20: Example outputs 1 under Setting 1.

Input:



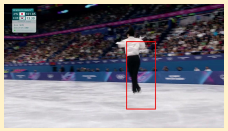
"text": "Changing feet Toe dig Quad toe did not work Awkward recovery.",
"description": "Changing feet Toe dig"

Ground-Truth:



```
"x": 51.14,
"y": 42.08,
"width": 22.36,
"height": 44.39,
"time": 80.4,
"description": "Changing feet Toe dig"
```

GPT5.6-sol



```
"bbox_2d": [555.0, 300.0, 690.0, 850.0],
"time": 81.0,
"description": "<Changing feet Toe dig> Quad toe did not work Awkward recovery."
```

Figure 21: Example outputs 2 under Setting 1.

C.2 SETTING2

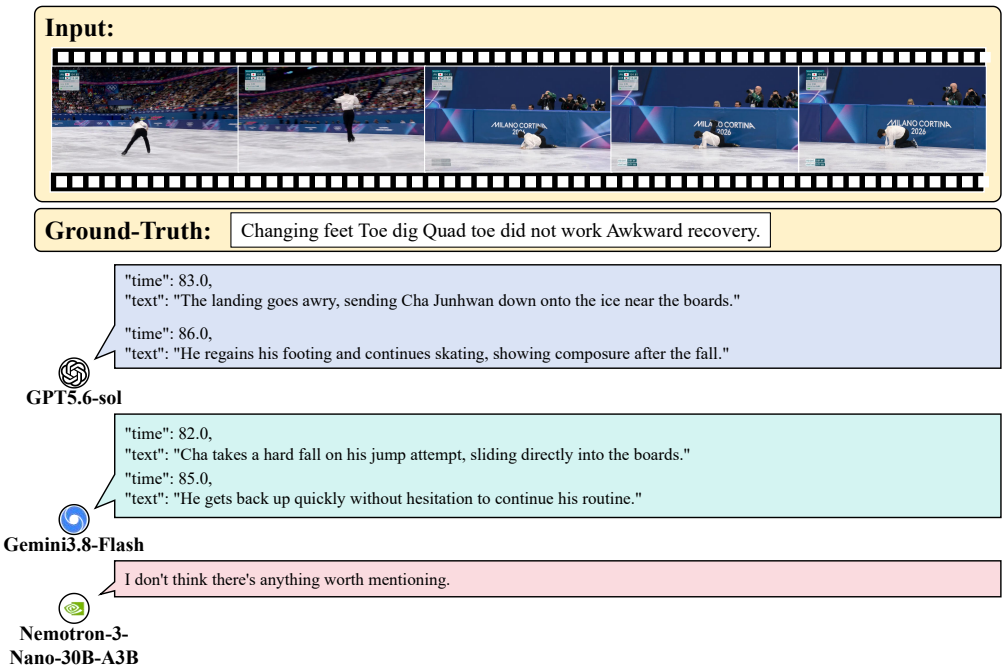


Figure 22: Example 2 outputs under Setting 1.

C.3 SETTING3

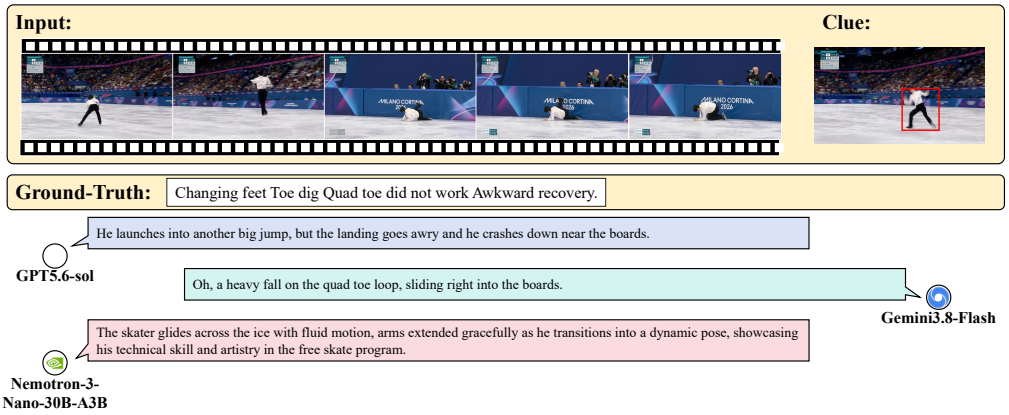


Figure 23: Example 3 outputs under Setting 1.