Do We Need Large VLMs for Spotting Soccer Actions?

Ritabrata Chakraborty* **Rajatsubhra Chakraborty**
Manipal University Jaipur **UNC–Charlotte**

Avijit Dasgupta **Sandeep Chaurasia**
IIIT Hyderabad **Manipal University Jaipur**

ritabrata.229301716@muj.manipal.edu

Abstract

Traditional video-based tasks like soccer action spotting rely heavily on visual inputs, often requiring complex and computationally expensive models to process dense video data. We propose a shift from this video-centric approach to a text-based task, making it lightweight and scalable by utilizing Large Language Models (LLMs) instead of Vision-Language Models (VLMs). We posit that expert commentary, which provides rich descriptions and contextual cues contains sufficient information to reliably spot key actions in a match. To demonstrate this, we employ a system of three LLMs acting as judges specializing in outcome, excitement, and tactics for spotting actions in soccer matches. Our experiments show that this language-centric approach performs effectively in detecting critical match events coming close to state-of-the-art video-based spotters while using zero video processing compute and similar amount of time to process the entire match.

1 Introduction

> Football is a game of mistakes. Whoever makes the fewest mistakes wins.
>
> Johan Cruyff

In the domain of video understanding (Nguyen et al., 2024), visual frames have traditionally been considered the best input for many tasks, including action spotting, event detection, and object recognition (Giancola et al., 2025, 2023; Fulari, 2018) . However, these methods often require significant computational resources to process and analyze the dense video data (Selva et al., 2023; Feichtenhofer et al., 2019). Despite the advancements in video models, such as convolutional neural networks (CNNs) (Karpathy et al., 2014) and vision transformers (ViTs), the need for high-resolution video inputs can be prohibitive in both training and deployment scenarios.

Action spotting (Seweryn et al., 2023), a core task in sports analytics, aims to identify key events within a video, such as goals, penalties, or substitutions, by analyzing the visual content. Manual methods by broadcasters were slow and took time in distribution (Merler et al., 2019).Traditional approaches (Shih, 2017) have relied on object detection and tracking techniques that require parsing every frame of the video to detect specific actions (Khan et al., 2018). These methods can be computationally expensive and often struggle with long sequences or multiple simultaneous events (Xu et al., 2025). In contrast, when considering the commentary, each moment in the match is often described in rich detail, including the action, the players involved, and the contextual relevance. The spoken word can provide a nuanced understanding of the match dynamics, capturing moments of excitement, controversy, and strategic importance that may not always be fully conveyed through visual data alone. This raises an interesting possibility: ***Can we leverage textual commentary as a primary input for action spotting, bypassing the need for video frames?***

We explore this question by proposing a text based action spotting pipeline using an LLM-as-a-judge setup, following (Zheng et al., 2023). We investigate whether expert commentary is enough for current LLMs to infer actions from, and if it is comparable to heavy video based action spotter VLMs. We also study the improvement in action spotting as time taken per match and the independence from video processing compute.To this end, we provide the following contributions:

- We redesign action spotting as a text based task as compared to a visual based task, utilising the Soccernet-Echoes dataset (Gautam et al., 2024).
- We design and implement a three-LLM system that judges the commentary based on out-

*Corresponding author

Proceedings of the IJCNLP-AACL 2025 Student Research Workshop, pages 59–65

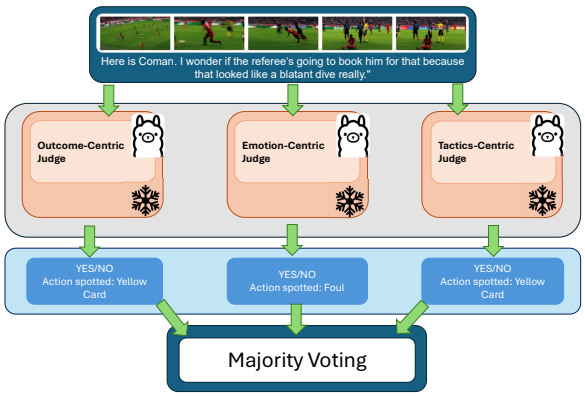

Figure 1: Our proposed LLM-based action spotting framework.

come , excitement, and tactics.

- We demonstrate that expert commentary, in many cases, provides comparable information for event detection compared to visual cues.
- We show that, by focusing on commentary alone, it is possible to detect key events reliably, highlighting the potential of language-centric models for sports analytics.

The rest of the paper is structured as follows. Section 2 discusses present literature around using text-based inputs for video tasks and action spotting in soccer matches. Section 3 explains our proposed framework in detail. Section 4 sheds light on the experimental setup and quantitative results. Finally we discuss some limitations in Section 6.

2 Related Work

Detailed Descriptions in Video-Based Tasks. In video-based understanding tasks, traditional models have primarily relied on visual features extracted from video frames to detect and classify events. However, recent research has begun exploring the use of fine-grained descriptions, specifically, textual information derived from transcriptions or commentary, to enhance performance in tasks like action spotting and event detection. Xie et al. (2019) demonstrated that integrating visual information with text can improve performance in action recognition tasks, as descriptive cues often convey context that is missed by raw visual data.In addition, Su et al. (2012) highlighted the utility of crowd-sourced commentary to aid in object detection tasks, which suggests that action spotting in dynamic environments, such as sports, could be enhanced by considering detailed narrative descriptions. Recent work shows that textual descriptions can carry action semantics the pixels miss and when transcribed reliably, can act as a compact surrogate for frames. For soccer specifically, dense, timestamped commentary corpora like SoccerNet-Caption (Mkhallati et al., 2023) and GOAL (Qi et al., 2023) establish the feasibility of commentary-anchored modeling, while MatchTime (Rao et al., 2024) highlights and fixes video-text misalignment—a key pain point for using commentary in downstream tasks. Robust automatic speech recognition (ASR) models such as Whisper (Radford et al., 2022) makes multi-accent, broadcast-noise transcripts viable at scale, strengthening the case for text-first pipelines.

Action Spotting in Soccer Videos. Action spotting in soccer has long relied on visual inputs, particularly tracking players and ball movements. However, recent developments in leveraging commentary and other textual sources for action detection have gained attention. Giannakopoulos et al. (2016) proposed a method that uses timestamped commentary as input to detect key moments in soccer, such as goals or penalties, demonstrating that textual data can complement traditional visual cues. Another approach by Andrews et al. (2024) used a multi-modal network that combines both video frames and textual commentary to detect key events in football matches. The SoccerNet benchmark (Deliege et al., 2021) formalized spotting as timestamp localization, driving a largely video-first literature . Classical baselines learn visual features and pool them temporally such as CALF (Cioppa et al., 2020) and NetVLAD++ temporal pooling (Giancola and Ghanem, 2021) . Subsequent models improved localization via stronger heads/sequence learning, including RMS-Net (Tomei et al., 2021) and compact E2E-Spot (Hong et al., 2022). Recent transformer systems such as ASTRA (Xarles et al., 2023) push tight-tolerance accuracy further and even add audio for non-visible cues. Broader universal efforts such as UniSoccer (Rao et al., 2025) argue for richer taxonomies and multi-task foundations that still place video at the center. These threads collectively set a strong video baseline for action spotting.

Despite these promising advancements, there

remains a gap in fully utilizing fine-grained commentary for video understanding tasks like action spotting, especially in the context of soccer. Existing methods either rely on computationally expensive visual cues or fail to achieve consistent performance with textual input alone.

3 Methodology

Large Language Model Judges. We use Llama 3.1 8B (Grattafiori et al., 2024) to instantiate three specialized judges that operate over a shared label space of the 17 SoccerNet-V2 classes and NO-ACTION. Each judge sees the same 10 s commentary window (5 s stride) but is prompted with a distinct *evidence lens* (Outcome, Tactics, Emotion). All three judges return a single class (or NO-ACTION) and confidence score. Judges are steered by a dedicated system prompt and 2–3 few-shot exemplars .

Outcome-centric Judge
Prioritizes refereeable outcomes (goal, penalty, yellow/red), explicit referee phrases.

Tactics-centric Judge
Emphasizes set-pieces and structure (corner, free-kick, substitution, formation/press).

Emotion-centric Judge
Uses rhetorical intensity and urgency to resolve ambiguous cases; conservative when negations appear ("over the bar", "flag is up").

Input: full English commentary for a 10 s window (5 s stride).
Output (per judge):

1. A **single label** in {17 SoccerNet-V2 classes} ∪ {NO-ACTION}.
2. A **confidence** in $[0, 1]$ (calibrated from model's self-score).

`Abstention is expressed as NO-ACTION; we use higher thresholds for the Emotion judge to avoid rhetorical over-triggering.`

Majority Voting System. Once each judge makes its decision, we aggregate the results using a majority voting mechanism. If at least two of the three judges agree on the presence of a relevant action, the action is considered "spottable" and is classified as an event worthy of attention. If the judges disagree, the action is not classified as relevant. This ensures that only the most unanimously recognized actions are selected.

Out-of-World Action Classification. In addition to the 17 predefined action classes, our system is designed to handle "out-of-world" actions—those that may be noteworthy but do not fall under any of the predefined classes. For instance, a player might execute a spectacular skill move or a controversial non-foul action, which can be exciting and relevant but doesn't match the typical goal or penalty. In this case, the judges are given the opportunity to classify the action as "out of world," providing a broader view of game dynamics that goes beyond standard categories.

Method	M	mAP (%)	Tight mAP (%)
CALF (Cioppa et al., 2020)	Video	49.7	–
RMS-Net (Tomei et al., 2021)	Video	63.49	28.83
FCMA (Zhou et al., 2021)	Video	73.77	47.05
E2E-Spot (RegNetY-200MF) (Hong et al., 2022)	Video	73.25	61.19
E2E-Spot (RegNetY-800MF) (Hong et al., 2022)	Video	74.05	61.82
ASTRA (Xarles et al., 2023)	Video	**78.09**	**66.82**
Random Text-Only (ours, baseline)	Text	**12.0**	**10.5**
LLM-Based (Ours)	**Text**	**64.5**	**60.8**

Table 1: mAP and tight mAP on SoccerNet-v2 for video- vs text-based pipelines. M = Modality (Video/Text).

4 Evaluation and Results

Setup. We evaluate on the SoccerNet-v2 test split using the SoccerNet-Echoes commentary (Gautam et al., 2024) as input. All commentary is in English. SoccerNet-Echoes provides timestamped transcriptions that are aligned to the underlying broadcast video using an automatic speech recognition (ASR) pipeline based on Whisper (Radford et al., 2022); we rely on these alignments without additional temporal adjustment. Events follow the 17-class SoccerNet taxonomy. Our system operates on 10 s windows (5 s stride) and uses three Llama 3.1 8B judges. Following SoccerNet (Deliege et al., 2021), we evaluate temporal localization using mean Average Precision (mAP) at multiple time tolerances. For a given tolerance δ (in seconds), a prediction of class c at time $\hat{\tau}$ is counted as correct if there exists a ground-truth event of class c at time τ such that $|\hat{\tau} - \tau| \leq \delta$. Let $\mathrm{AP}(\delta)$ denote the Average Precision over all events at tolerance δ. We then define

$$\mathrm{mAP} = \frac{1}{|\Delta_{\text{loose}}|} \sum_{\delta \in \Delta_{\text{loose}}} \mathrm{AP}(\delta), \qquad (1)$$

$$\text{Tight mAP} = \frac{1}{|\Delta_{\text{tight}}|} \sum_{\delta \in \Delta_{\text{tight}}} \mathrm{AP}(\delta), \qquad (2)$$

where $\Delta_{\text{loose}} = \{5, 10, 15, \ldots, 60\}$ s and $\Delta_{\text{tight}} = \{1, 2, 3, 4, 5\}$ s. We report both aggregates in Tables 1 and 2. For efficiency (Table 2) we normalize video compute to a **90 min match at 2 FPS** (10,800 frames) and report backbone FLOPs/frame

Method	Input	FLOPs/frame (GF)	Total FLOPs (TF)	Time/frame (ms)	Time/match (sec)
RegNetY-200MF (E2E-Spot)	Video	0.20	2.16	0.3	3.24
RegNetY-800MF (E2E-Spot)	Video	0.80	8.64	0.9	9.72
ResNet-152 (baseline feats)	Video	11.5	124.2	1.8	19.44
R(2+1)D (3D CNN)	Video	–	–	11.0	118.8
Ours: LLM (text-only)	**Text**	**–**	**–**	**–**	146.5

Table 2: Efficiency comparison on FLOPs and wall-clock time required for a full match evaluation. Video times are reported for GPU backbone-only inference on an A5000 (excluding video decoding and post-processing). "Ours" reports full end-to-end CPU time; although the wall-clock is larger, our method incurs zero video FLOPs and does not require a GPU. Here, FLOPs denotes floating point operations, and GF and TF correspond to 10^9 and 10^{12} FLOPs, respectively.

and measured per-frame time on an A5000 from prior work (Hong et al., 2022). For our text-only system we do not process any frames and report the end-to-end wall-clock time to process a full 90 min match on a single commodity CPU (16-core, 32 GB RAM).

Textual Random Baseline. To establish a strict lower bound we create a commentary-anchored randomness baseline that predicts actions without reading the text. For each commentary sentence $s_k = (\tau_k, \ell_k, \text{text}_k)$ in a half we sample a Bernoulli coin for every action class c with probability equal to that class's empirical commentary prior p_c (estimated on the train split). If the coin succeeds we emit a pseudo detection $(\tau_k, c, 0.5)$; overlapping detections of the same class within 2δ (δ=10 s) are merged by keeping the earliest. Here π_k denotes the prior frequency of class k in the training split, l_k is the unnormalized logit score predicted for class k, and text_k is the k-th commentary sentence or window. This design respects the real timestamp distribution yet ignores all lexical information, yielding the hardest chance-level floor against which any text-aware model must improve.

Main results. Table 1 shows that our text-only system achieves **64.5 mAP** and **60.8 Tight**, substantially outperforming the *Random Text-Only* baseline (12.0 / 10.5) and approaching recent video methods despite using no visual frames. Relative to strong video pipelines, we are close on the tight metric (**60.8** vs **61.82** for E2E-Spot RegNetY-800MF; **60.8** vs **66.82** for ASTRA), while trailing more on loose mAP (**64.5** vs **74.05** and **78.09**). Compared to RMS-Net, our tight score is more than **2**× higher (60.8 vs 28.83) and our loose mAP is competitive (64.5 vs 63.49). The pattern aligns with the nature of commentary: explicitly lexicalized, refereeable outcomes (goals, penalties, bookings, substitutions) are well localized in time, benefitting Tight mAP; at larger tolerances we remain intentionally conservative via abstention, trading some recall for precision.

Efficiency ablation. Table 2 compares per–90 min match compute and explains our savings. Video pipelines pay a cost that scales with the number of *visual tokens*; text scales with *text tokens*. Let F be frames per match and P the patch tokens per frame (ViT-style). Then visual-tokens $= F \times P$ and text-tokens $= N_t$. At 2 FPS, F=10,800. For ViT-B/16 at 224^2, $P=(224/16)^2=196$, so $F \times P \approx 2.12 \times 10^6$ visual tokens/match, whereas ASR produces only $N_t = \mathcal{O}(10^4)$ text tokens—two orders of magnitude fewer. Even with CNNs (no explicit patches), the effective per-frame compute (GFLOPs/frame) still scales with F and dominates.

At 2 FPS, published video backbones span 2.16–124.2 TFLOPs per match and 0.3–1.8 ms per frame on an A5000 (3.24–19.44 s per match; a 3D CNN is 118.8 s). Our pipeline performs no video feature extraction (zero video FLOPs) and instead scales with N_t and LLM tokens/s. On CPU, our measured end-to-end time for a full match is 146.5 s (2.44 min), removing the dominant frame-processing term and any GPU requirement.

Discussion. (1) **Tight localization from text.** When outcomes are spoken ("penalty given", "booked", "and it's in"), the language signal is temporally sharp, explaining our proximity to video SOTA on Tight mAP. (2) **Loose-gap sources.** Non-verbal micro-events and terse restarts are under-described in commentary, which hurts loose recall and favors video. (3) **Design effects.** Confidence thresholds and majority voting suppress rhetorical false positives (near-misses), improving precision; temporal NMS converts overlapping window votes

into a single timestamp per event. (4) **Compute and deployment.** Zero-frame processing plus competitive Tight mAP make the approach attractive for CPU-scale batch processing (clubs/broadcasters) and for low-cost inference at volume. In summary, the main advantages of the text-based formulation are: (i) no need to store or process video frames, (ii) CPU-only inference with predictable scaling in the number of commentary tokens, and (iii) strong performance on refereeable, explicitly verbalized events (goals, penalties, cards, substitutions). The main drawbacks are: (i) a hard dependence on commentary coverage and timing, (ii) limited access to visual cues that are never spoken aloud (e.g., off-ball incidents or subtle shape changes), and (iii) potential lack of generalization to matches or leagues with minimal or low-quality commentary. We return to this trade-off in Section 5.

5 Conclusion

In this paper we asked whether large vision–language models are necessary to spot soccer actions when high-quality expert commentary is available. By reformulating action spotting as a purely language-centric task and applying a three-judge LLM ensemble to 10 s commentary windows, we show that text-only spotting can approach the performance of recent video-based systems: our method achieves 64.5 mAP and 60.8 Tight mAP on SoccerNet-v2, reaching 83%–96% of ASTRA's video-based performance while using zero video processing compute.

Our results suggest a nuanced answer to the title question. When dense, time-aligned commentary is present—as in professional broadcasts with experienced commentators—we do *not* strictly need VLMs for many refereeable events (goals, penalties, cards, substitutions). In this regime, language carries most of the necessary semantics and can be processed on commodity CPUs without maintaining or streaming video frames. However, when commentary is sparse, noisy, delayed, or entirely absent, or when the task depends on fine-grained visual cues that commentators do not verbalize (e.g., subtle tactical shapes, off-ball incidents, or crowd reactions), vision-based models remain indispensable.

Looking forward, we see text-only spotting as a strong and complementary baseline rather than a replacement for VLMs. A promising direction is to build multimodal pipelines where commentary provides a high-level prior over candidate events, and lightweight video modules are invoked only when the text is ambiguous or inconsistent with the visual evidence. Such hybrids could retain most of the efficiency gains of our language-centric design while recovering the visual coverage needed in more challenging or low-commentary scenarios.

6 Limitations

While our framework shows promising results, there are several limitations to consider. First, the performance of our system is heavily dependent on the quality of the commentary and transcription. Inaccurate or incomplete commentary can hinder the ability of our judges to correctly identify action-worthy events, leading to lower accuracy in the action spotting task. Similarly, the quality of transcription performed by Whisper plays a critical role. Errors in the transcription process can result in incorrect words or misplaced timestamps, directly affecting the action spotting metrics, including mean Average Precision (mAP). These transcription errors could affect the reliability of the timestamped actions and ultimately influence the results of the semantic judging. A second cost that we do not explicitly quantify in Tables 1–2 is automatic speech recognition. In our pipeline the Whisper-based ASR step is run once per match to produce commentary transcripts and can be executed offline or cached for reuse across downstream tasks. Nevertheless, ASR incurs its own compute and latency costs that broadcasters and practitioners must account for in an end-to-end system design; a fully fair comparison to video-only pipelines should include this term, which we leave for future work. Additionally, our framework assumes that the provided commentary is sufficiently detailed and relevant for the action spotting task. In cases where the commentary lacks context or important details, the system's performance may degrade. We aim to address this in our future work.

References

Peter Andrews, Oda Elise Nordberg, Stephanie Zubicueta Portales, Njål Borch, Frode Guribye, Kazuyuki Fujita, and Morten Fjeld. 2024. Aicommentator: A multimodal conversational agent for embedded visualization in football viewing. In *Proceedings of the 29th International Conference on Intelligent User Interfaces*, pages 14–34.

Anthony Cioppa, Adrien Deliège, Silvio Giancola,

Bernard Ghanem, Marc Van Droogenbroeck, Rikke Gade, and Thomas B. Moeslund. 2020. A context-aware loss function for action spotting in soccer videos. *Preprint*, arXiv:1912.01326.

Adrien Deliege, 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. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4508–4519.

Christoph Feichtenhofer, Haoqi Fan, Jitendra Malik, and Kaiming He. 2019. Slowfast networks for video recognition. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 6202–6211.

Sunit Fulari. 2018. A survey on motion models used for object detection in videos. In *2018 2nd International Conference on Trends in Electronics and Informatics (ICOEI)*, pages 348–353.

Sushant Gautam, Mehdi Houshmand Sarkhoosh, Jan Held, Cise Midoglu, Anthony Cioppa, Silvio Giancola, Vajira Thambawita, Michael A Riegler, Pal Halvorsen, and Mubarak Shah. 2024. Soccernet-echoes: A soccer game audio commentary dataset. In *2024 International Symposium on Multimedia (ISM)*, pages 71–78. IEEE.

Silvio Giancola, Anthony Cioppa, Julia Georgieva, Johsan Billingham, Andreas Serner, Kerry Peek, Bernard Ghanem, and Marc Van Droogenbroeck. 2023. Towards active learning for action spotting in association football videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pages 5098–5108.

Silvio Giancola, Anthony Cioppa, Bernard Ghanem, and Marc Van Droogenbroeck. 2025. *Deep Learning for Action Spotting in Association Football Videos*, page 427–459. WORLD SCIENTIFIC.

Silvio Giancola and Bernard Ghanem. 2021. Temporally-aware feature pooling for action spotting in soccer broadcasts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4490–4499.

Theodoros Giannakopoulos, Anastasios Tsoumakas, and Ioannis Vlahavas. 2016. Soccer action spotting with timestamped commentary. In *Proceedings of the 17th International Conference on Artificial Intelligence: Methodology, Systems, and Applications, AIMSA 2016, Varna, Bulgaria, September 7-10, 2016*, pages 309–318. Springer.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

James Hong, Haotian Zhang, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. 2022. Spotting temporally precise, fine-grained events in video. In *European Conference on Computer Vision*, pages 33–51. Springer.

Andrej Karpathy, George Toderici, Sanketh Shetty, Thomas Leung, Rahul Sukthankar, and Li Fei-Fei. 2014. Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition*, pages 1725–1732.

Abdullah Khan, Beatrice Lazzerini, Gaetano Calabrese, Luciano Serafini, and 1 others. 2018. Soccer event detection. *Computer Science & Information Technology*, pages 119–129.

Michele Merler, Khoi-Nguyen C. Mac, Dhiraj Joshi, Quoc-Bao Nguyen, Stephen Hammer, John Kent, Jinjun Xiong, Minh N. Do, John R. Smith, and Rogerio Schmidt Feris. 2019. Automatic curation of sports highlights using multimodal excitement features. *IEEE Transactions on Multimedia*, 21(5):1147–1160.

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

Thong Nguyen, Yi Bin, Junbin Xiao, Leigang Qu, Yicong Li, Jay Zhangjie Wu, Cong-Duy Nguyen, See-Kiong Ng, and Anh Tuan Luu. 2024. Video-language understanding: A survey from model architecture, model training, and data perspectives. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 3636–3657, Bangkok, Thailand. Association for Computational Linguistics.

Ji Qi, Jifan Yu, Teng Tu, Kunyu Gao, Yifan Xu, Xinyu Guan, Xiaozhi Wang, Bin Xu, Lei Hou, Juanzi Li, and 1 others. 2023. 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*, pages 5391–5395.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *Preprint*, arXiv:2212.04356.

Jiayuan Rao, Haoning Wu, Hao Jiang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2025. Towards universal soccer video understanding. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 8384–8394.

Jiayuan Rao, Haoning Wu, Chang Liu, Yanfeng Wang, and Weidi Xie. 2024. MatchTime: Towards automatic soccer game commentary generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 1671–1685, Miami, Florida, USA. Association for Computational Linguistics.

Javier Selva, Anders S. Johansen, Sergio Escalera, Kamal Nasrollahi, Thomas B. Moeslund, and Albert Clapés. 2023. Video transformers: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(11):12922–12943.

Karolina Seweryn, Anna Wróblewska, and Szymon Łukasik. 2023. Survey of action recognition, spotting and spatio-temporal localization in soccer–current trends and research perspectives. *arXiv preprint arXiv:2309.12067*.

Huang-Chia Shih. 2017. A survey of content-aware video analysis for sports. *IEEE Transactions on circuits and systems for video technology*, 28(5):1212–1231.

Hao Su, Jia Deng, and Li Fei-Fei. 2012. Crowdsourcing annotations for visual object detection. *HCOMP@ AAAI*, 1.

Matteo Tomei, Lorenzo Baraldi, Simone Calderara, Simone Bronzin, and Rita Cucchiara. 2021. Rms-net: Regression and masking for soccer event spotting. In *2020 25th International Conference on Pattern Recognition (ICPR)*, pages 7699–7706. IEEE.

Artur Xarles, Sergio Escalera, Thomas B Moeslund, and Albert Clapés. 2023. Astra: An action spotting transformer for soccer videos. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, pages 93–102.

Ning Xie, Farley Lai, Derek Doran, and Asim Kadav. 2019. Visual entailment: A novel task for fine-grained image understanding. *arXiv preprint arXiv:1901.06706*.

Hao Xu, Arbind Agrahari Baniya, Sam Well, Mohamed Reda Bouadjenek, Richard Dazeley, and Sunil Aryal. 2025. Action spotting and precise event detection in sports: Datasets, methods, and challenges. *Preprint*, arXiv:2505.03991.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

Xin Zhou, Le Kang, Zhiyu Cheng, Bo He, and Jingyu Xin. 2021. Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection. *arXiv preprint arXiv:2106.14447*.