Towards Automatic Soccer Commentary Generation with Knowledge-Enhanced Visual Reasoning

Zeyu Jin[1,2], Xiaoyu Qin[1,*], Songtao Zhou[1], Kaifeng Yun[1], Jia Jia[1,2,*]
[1]Department of Computer Science and Technology, Tsinghua University, [2]BNRist, Tsinghua University
{jinzeyu23, zst24, ykf21}@mails.tsinghua.edu.cn, {xyqin, jjia}@tsinghua.edu.cn

***Abstract*—Soccer commentary plays a crucial role in enhancing the soccer game viewing experience for audiences. Previous studies in automatic soccer commentary generation typically adopt an end-to-end method to generate anonymous *live text commentary*. Such generated commentary is insufficient in the context of real-world *live televised commentary*, as it contains anonymous entities, context-dependent errors and lacks statistical insights of the game events. To bridge the gap, we propose GAMESIGHT, *a two-stage model to address soccer commentary generation as a knowledge-enhanced visual reasoning task*, enabling live-televised-like knowledgeable commentary with accurate reference to entities (players and teams). GAMESIGHT starts by performing visual reasoning to align anonymous entities with fine-grained visual and contextual analysis. Subsequently, the entity-aligned commentary is refined with knowledge by incorporating external historical statistics and iteratively updated internal game state information. Consequently, GAMESIGHT improves the player alignment accuracy by 18.5% on SN-Caption-test-align dataset compared to Gemini 2.5-pro. Combined with further knowledge enhancement, GAMESIGHT outperforms in segment-level accuracy and commentary quality, as well as game-level contextual relevance and structural composition. We believe that our work paves the way for a more informative and engaging human-centric experience with the AI sports application. Demo page: https://gamesight2025.github.io/gamesight2025.**

I. INTRODUCTION

The soccer industry holds a significant position in the global sports market, with a fan base of over five billion people [1]. As one of the key elements of the viewing experience, soccer commentary is of vital importance in enhancing fan engagement and delivering captivating information. This has sparked considerable interest in automatic soccer commentary generation [2], [3]. Given a segment of soccer game video, end-to-end soccer commentary generation models, e.g., MatchVision [4] and SoccerComment [5], primarily focus on producing descriptions in the style of anonymous *live text commentary*. However, they fall short when it comes to replicating the real-world *live televised commentary*.

Professional live televised commentary is characterized by several key elements to captivate the audience: **(1) Accurate reference to entities [6].** By aligning names of players and teams in the sportscasting footage and the ongoing events, game commentators help the audience to understand the unfolding action. **(2) Awareness of the internal game context.** Commentators also integrate the dynamic match context to offer a comprehensive interpretation of current game state.

*Corresponding authors: Xiaoyu Qin and Jia Jia.

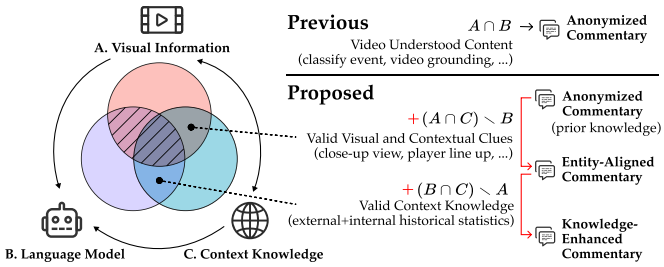

Fig. 1: Overview of the proposed GAMESIGHT. It incorporates visual and contextual information to generate soccer commentary that is entity-aligned and knowledge-enhanced.

The commentary paragraph should flow continuously, rather than being fragmented, to maintain a cohesive narrative. **(3) Utilization of external statistics**. Commentary should consist of not only description, but also explanation and comment [7]. Combining historical statistical knowledge serves as the foundation of these broader perspective and deeper insights [8].

Previous soccer commentary models fail to meet the above from the following aspects. **Prob. 1: Absence of entity grounding.** Previous models generate lines with anonymized player and team labels as placeholders [2], incompetent for aligning the specific players involved in the events. **Prob. 2: Context-dependent errors.** These models are unaware of broader context as they aim solely at increasing the text similarity in video captioning. This leads to context-dependent errors in critical game information, such as incorrect score announcements following goal events. **Prob. 3: Lack of statistical insight.** Live text commentary inherently focus on descriptive commentary due to the absence of video, which differs from live televised commentary. As a result, previous models lack the incorporation of statistical knowledge.

To tackle the three key discrepancies between end-to-end soccer commentary models and human-centric televised commentary, we propose GAMESIGHT, a two-stage model that addresses soccer commentary generation as a knowledge-enhanced visual reasoning task. We move beyond the limited end-to-end captioning by decoupling the task into visual entity grounding and knowledge-driven refinement. This design mimics the cognitive process of human commentators: first identifying the participants through visual and contextual cues, and then enriching the narrative with deep domain knowledge, as illustrated in Fig. 1. By integrating both internal

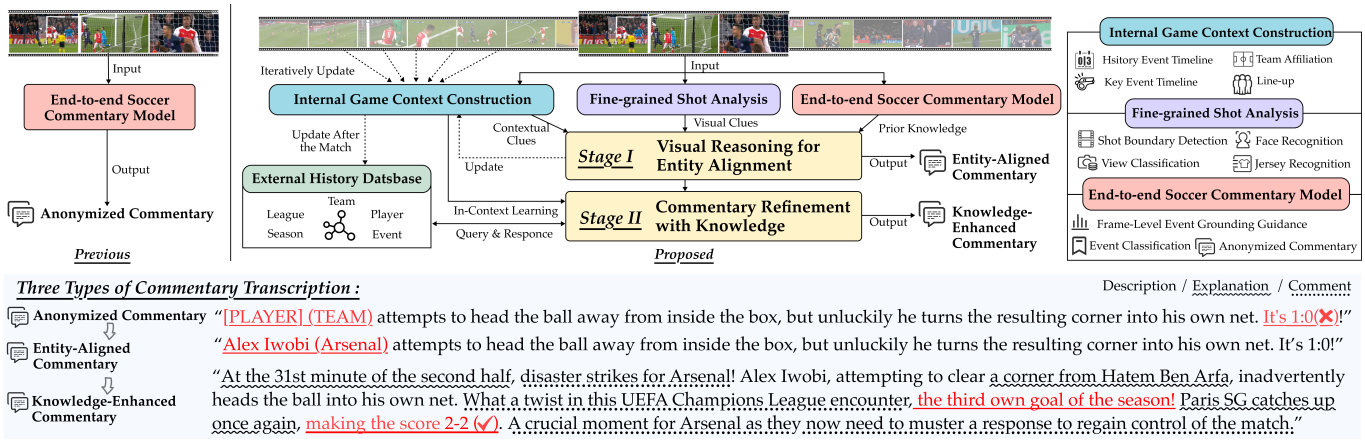

Fig. 2: Unlike previous end-to-end models, GAMESIGHT is a two-stage system that addresses soccer commentary generation as a knowledge-enhanced visual reasoning task. It revises and enriches the anonymized commentary with *Stage I*: visual reasoning for entity alignment, and *Stage II*: commentary refinement with knowledge.

game context and external historical statistics, GAMESIGHT produces commentary that is not only visually accurate but also contextually profound.

In summary, our contributions are three-fold:

- We introduce soccer commentary generation as a knowledge-enhanced visual reasoning task, bridging the gap between automated live-text like commentary and human-centric live televised like commentary.
- We propose GAMESIGHT, a two-stage system that leverages fine-grained visual and contextual information to conduct entity alignment and knowledge enhancement for commentary.
- Experiments show that GAMESIGHT outperforms in both segment-level and game-level to generate accurate, informative and engaging commentary, showcasing the capability of AI-based method to enrich the viewing experience in sports analysis.

II. PROBLEM FORMULATION

To bridge the gap, we formulate soccer commentary generation as a **knowledge-enhanced visual reasoning task**. Given a soccer video segment $\mathcal{S}$ and its corresponding initial anonymized commentary C_A (typically generated by an end-to-end model), our goal is to produce a final commentary C_{KE} that is both entity-accurate and contextually insightful.

A. The Two-Stage Design Rationale

Unlike previous end-to-end approaches, we decouple the task into two specialized stages (as shown in Fig. 2) to address the three problems:

Stage I: Align Entity with Visual Reasoning (Prob. 1). As shown in the pre-experiment (Sec. III-A), human rely on a rich set of contextual and visual cues to understand the event-related players besides the player tracking [9] in long view scenes. Inspired by this insight, Stage I is trained with *supervised fine-tuning* (SFT) and *group relative policy optimization* (GRPO) to combine internal game context and fine-grained shot analysis to transform C_A into entity-aligned commentary C_{EA} (Sec. III-E).

Stage II: Commentary Refinement with Knowledge (Prob. 2 & 3). An analysis of live televised commentators' practices (Sec. IV-A) reveals their frequent and flexible application of relevant knowledge. In line with this behavior, we develop a soccer *knowledge-augmented generation* (KAG) system and an iteratively updated database of internal game context knowledge in Stage II to generate C_{KE}.

B. Formal Objectives

Formally, the pipeline is modeled as two successive stages:

Stage I (Sec III): $f_{align}(\mathcal{S}, C_A, Context) \rightarrow C_{EA}$. This function focuses on maximizing the grounding accuracy of C_{EA} relative to the ground-truth players in $\mathcal{S}$. The $Context$ includes contextual clues C_{GS} (Sec. III-B), visual clues $Scene$ (Sec. III-C), and the prior knowledge guidance $\mathcal{W}$ (Sec. III-D).

Stage II (Sec IV): $f_{enrich}(C_{EA}, \mathcal{K}_{ext}, \mathcal{K}_{int}) \rightarrow C_{KE}$. This function focuses on generating $\mathcal{C}_{KE}$ with not only statistically accurate and contextually related commentary, but also insightful explanations and comments in a training-free manner, where $\mathcal{K}_{ext}$ (Sec. IV-B) and $\mathcal{K}_{int}$ (Sec. IV-C) denote external statistics and internal match history, respectively.

III. STAGE I: ALIGN ENTITY WITH VISUAL REASONING

A. Empirical Study

Given a soccer video segment and a piece of anonymized commentary, we develop a "Player Guessing" game and recruit three expert soccer commentators to take part in the study on five matches. During the game, once the participants get a correct answer, they are request to recall the clues they rely on to pinpoint the player. Consequently, human participants have an accuracy of **96.3%**. As illustrated in Fig. 4, the correct identifications are made through various cues. Notably, **84.5%** of the correct identifications were done by more than just long

Fig. 3: Soccer game view shots distribution in time length and view shots examples.

view player tracking, indicating the necessity of involving fine-grained visual and contextual information in player alignment.

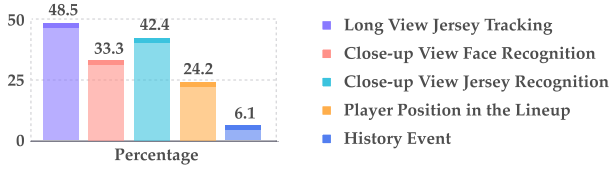

Fig. 4: Human reliance for the "Player Guessing" game.

B. Internal Game Context Construction

The context game state C_{GS} iteratively updates throughout the game. It mainly includes the temporal team line-up $\mathcal{T}$, $KeyEvent$ timeline, and $HistoryEvent$ timeline. For a match between the home team $Team_h$ ($Color_h$ jersey) and away team $Team_a$ ($Color_a$ jersey), the construction of each component is elaborated in Appendix Sec. D-A.

C. Fine-grained Shot Analysis

As shown in Fig. 3, shot views in a soccer video are usually categorized into four classes: long view, medium view, close-up view, and out-of-field view [10]. In the soccer game sportscasting footage, players appearing in the preceding or following close-up and medium views are more likely to be related to events recently happened on the field. Therefore, they provide potentially identity information to pinpoint the anonymized player. For the purpose of fine-grained details extraction, "medium view" and "close-up view" are collectively referred to as "close view" in the following sections. We apply several processing steps to capture the fine-grained visual details $Scene$ of the input video, including (1) Shot boundary detection with view classification, (2) Player face recognition, (3) Player jersey recognition and (4) Team affiliation detection. See Appendix Sec.D-B for details in each procedure.

D. Frame-Level Event Grounding Guidance

A 30-second video segment typically contains far more events than the described one, as soccer video clips are not trimmed at the event level but centered around the ground-truth timestamp. For example, a corner kick may seamlessly shift into open play. Therefore, identifying the correct event within this fixed-length segment is essential.

However, foundational video-LMMs often struggle to capture complex multi-player dynamics in the soccer domain. To address this, we leverage the state-of-the-art soccer commentary model, MatchVision [11], to obtain prior knowledge between videos and commentary events. Since the precise commentary response can only be generalized from the accurately grounded video segments, we obtain a frame-wise vector $\mathcal{W}$ from the Q-former's cross-attention layers of MatchVision to represent the relevance of each video frame to the generated narration (Details provided in Appendix Sec.D-C). It can be seen as the arousal of each frame in generating the commentary text, which serves as a frame-level event grounding guidance.

E. Instruction Tuning with video-LMM

Given the video $\mathcal{S}$ and anonymized commentary $\mathcal{C}_A$, with $Context = \{C_{GS}(t), Scene(t), \mathcal{W}(t)\}$ as prompt, we leverage Video-LMM's visual understanding and reasoning capability to infer the correct answer $Player_g$ and his team affiliation $Team_g$ (example shown in Appendix Fig. 1). Due to the suboptimal performance of the base video-LMM model, we employs two kinds of fine-tuning strategies to fine-tune the base model. Firstly, we adopt SFT to establish task-specific grounding. We further implement the GRPO approach to further enhance the model's compositional reasoning capability and reward alignment in this single choice problem.

To prepare the training data for this visual reasoning task, we use the answer-guided reasoning strategy to decompose the problem with *Chain-of-Thought* reasoning. Through this process, we generate a train set guided by the correct answer reasoning procedure to bootstrap reasoning with reasoning [12]. For data augmentation, we query about the [Team] and [Player] separately with different CoT strategies (See prompt details in Appendix Sec.D-J).

IV. Stage II: Refine Commentary with Knowledge

A. Knowledge Application in Sportscasting

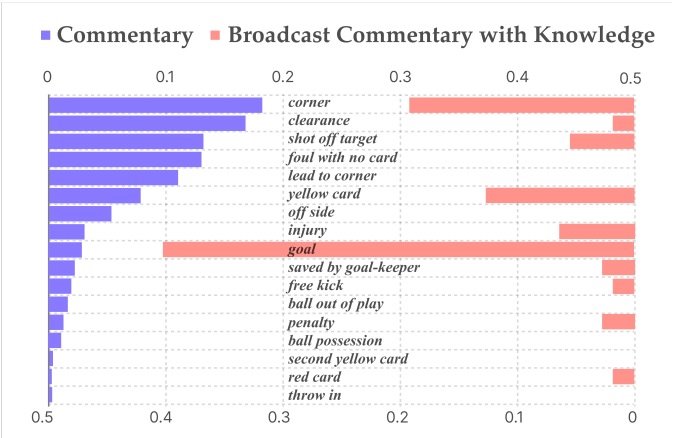

Fig. 5: Distribution of commentary at event level.

Live televised commentary exhibits diverse styles due to factors like the commentators' standpoint, oral habits, personality, and cultural background. While it is hard to reconcile these differences, human commentators share a mutual preference for incorporating contextual knowledge to offer explanations and comments on the visual scenes.

We compared the knowledge injection in live text commentary and audio transcriptions of live text commentary [5]. Inspired by FLARE [13], knowledge is extracted by LLM through paired explicit query formulations. We prompt GPT-4o to identify the expert knowledge embedded within the commentary. Additionally, we employ a self-asking mechanism [14] to generate constrained questions grounded in the extracted knowledge. It is found that 15.02% of the live televised commentary contains various knowledge, whereas only 3.16% of the live text commentary includes knowledge and is mostly limited to game score updates. The distribution is illustrated in Fig. 5. The major events where human commentators tend to reference knowledge include $Goal$, $Corner$, and $Card$. Furthermore, based on the source of the information, knowledge are categorized into **external knowledge**, which rely on the historical statistics from other games, and **internal knowledge**, which refers to the static background information and dynamic updated events from the current game.

B. External Soccer KAG System Modification

The soccer KAG system, built upon SoccerRAG [15], is designed for acquiring external statistics. To better emulate human commentator behavior, we propose several key enhancements. Firstly, we expanded the set of query exemplars in high-frequency commentary-related questions to enhance accuracy in the transition between natural language and SQL queries. We also developed augmented data schema to add support in finer-grained event-level queries, such as own-goal and penalty for players and teams. Furthermore, we designed strict temporal constraints within the query generation schema to ensure future matches excluded from retrieval. It maintains the temporal integrity of the database and prevents data leakage during inference.

C. Internal Game Context Tracking

The internal game context module tracks in-game information essential for commentary generation. Its construction follows an approach similar to Sec. III-B, with adjustments guided by the human commentators' reference habits (Sec. IV-A). Detailed background about the mentioned player, such as nationality and height, are provided in this module. Fine-grained information such as the goal scorer, assisting players, and the specific method of goal achievement (e.g., penalty, header, or own goal) are also captured.

D. Commentary Refinement with LLM

With the stage I commentary $\mathcal{C}_{EA}$ describing the basic current event, we use GPT-4o to generate questions related to external statistics. The responses from the KAG system are then double checked with LLM to discard the invalid answers with repetition or incorrect temporal range. In addition, internal game context knowledge, such as the key timeline of the mentioned event and detailed player information, are explicitly prompted for the model to quote. This process enhances the accuracy and insight of the commentary, making it more aligned with live televised commentary.

V. Experiments

In this section, we first introduce the experimental setup in Sec. V-A, then unfold two series of experiments for each stage of GameSight. As to stage I, Sec. V-B shows experiments on the accuracy of entity alignment. For stage II, we evaluate the performance of GameSight refined commentary in Sec. V-C from segment-level: baseline comparison, knowledge accuracy, and game-level: structural similarity, sentiment polarity and contextual relevance.

A. Experimental Setup

Stage I. We use the test set from SN-Caption-test-align (49 games, 2,305 segments) and train set from MatchTime (419 games, 18,826 segments). Zero-shot experiments on video-LMMs are conducted with video fps=1 and video quality in 720p (1,280×720 pixels) for the clearest visual details. Instruction-tuning experiments are conducted on 4 × H100 (80G) GPUs, with fps=1 and video quality in 180p (320×180 pixels) due to the resource limit. We use Qwen2.5-VL-7B-Instruct as the backbone model. SFT experiments are trained for 2 epochs with a learning rate of 1e-4. LoRA [16] is adopted with a rank of 8. GRPO experiments are trained each for 600 steps with binary reward. Learning rate is 1e-6 and β coefficient is 0.1. Following [17], we only utilize video, query, and the ground truth answer for training.

We report $Player$ alignment accuracy, and derive $Team$ alignment accuracy based on the predicted player's team. $Player_c$ and $Team_c$ denote settings where the correct team affiliation is given, reducing the candidate search space.

Stage II. As to segment-level commentary generation, we use SN-Caption-test-align (live text) and Goal benchmark [6] (live televised) as two versions of ground truth. The state-of-the-art soccer commentary models MatchVoice [4] and MatchVision [11] are chosen as the baselines. We employ BLEU [18], CIDER [19], METEOR [20] and ROUGE [21] as evaluation metrics. All commentary are anonymized for a fair comparison. For the internal knowledge accuracy evaluation, we conduct knowledge refinement using ICL knowledge only, and extract the referred internal statistics using the same prompt as Sec. IV-A. Catering to external knowledge accuracy, we manually tailor a set of 500 statistical questions reflecting human commentators' reference habits. Detailed samples are elaborated in Appendix F-A.2.

For game-level evaluation, we consider two baselines: the game-wise concatenated live text commentary [22], and the whole 90 minutes sportscast's ASR transcription [23]. In quantitative analysis, we adopt human evaluation, sentiment polarity, and the *Coh-Metrix* [24] in discourse analysis to evaluate the coherence and contextual relevance. For the structural

composition of commentary, sport communication experts [7] defines three logical categories: description, explanation, and comment, with the criteria that description in live televised commentary should not exceed 50%. We use GPT-3.5-turbo to conduct the classification.

B. Accuracy of Entity Alignment

Exp. 1: Zero-shot accuracy test of video-LMMs and LLMs. We select Qwen2.5-VL-7B-Instruct [25], InternVL3 [26], VideoLLaMA3-7B [27], and LLaVA-OneVision [28] as the representative of open-sourced video-LMMs, Gemini 2.5-pro [29] for close-sourced video-LMM, GPT-4o [30], GPT-o1-preview [31] and DeepSeek-R1 [32] for close-sourced LLMs. We additionally use the top-1/3 accuracy for player alignment. Tab. I demonstrates the zero-shot abilities of all kinds of video-LMMs and LLMs in the player alignment reasoning task. Although the contextual and visual clues can be organized in text format, LLMs performance falls short of the open-sourced video-LMMs, indicating the importance of involving visual modality to further observe the related scene. While Qwen2.5-VL lags behind close-sourced video-LMM in performance, it outperforms open-sourced models and most LLMs, which serves as the superior backbone for further fine-tuning.

TABLE I: Zero-shot performance of video-LMMs and LLMs. Bolded and underlined for the 1st and 2nd largest value.

Model	$Player@1$	@3	$Team$	$Player@1_c$	$@3_c$	$Team_c$
Qwen2.5-VL	<u>35.4</u>	<u>39.3</u>	70.9	<u>43.8</u>	48.6	91.4
InternVL3	26.3	39.1	<u>72.2</u>	30.8	47.4	92.0
VideoLLaMA3	17.2	27.5	63.7	19.4	36.4	83.7
LLaVA-OV	18.7	18.9	66.7	21.7	21.9	87.9
Gemini 2.5-pro	**52.6**	**55.9**	**71.2**	**61.3**	<u>62.4</u>	<u>96.9</u>
GPT-4o	21.2	24.0	49.5	37.7	45.0	98.0
GPT-o1	30.7	37.9	53.9	40.8	<u>62.8</u>	**98.2**
DeepSeek-R1	23.8	28.6	42.9	35.9	**63.2**	97.6

Exp. 2: Instruction tuning with video-LMM. Tab. II validates the effectiveness of the proposed fine-tuning strategy. Introducing SFT markedly improves $Player$ accuracy by 27.3% compared with base model, demonstrating strong task grounding with SFT. GRPO alone is less effective than SFT in boosting the performance, but it still shows that structured reasoning benefits from reward-driven alignment with a modest improvement. The combination of both SFT and GRPO leads to the best overall performance, with substantial gains in all categories besides $Team_c$. The synergy between SFT and GRPO proves crucial for addressing the multifaceted challenges of visual reasoning in video-LMMs.

TABLE II: Experiment results of fine-tuning on Qwen2.5-VL.

Training Strategy	$Player\uparrow$	$Team\uparrow$	$Player_c\uparrow$	$Team_c\uparrow$
Qwen2.5-VL #1	35.9	69.6	43.9	92.5
#1 + SFT	<u>63.2</u>	<u>81.7</u>	<u>67.0</u>	<u>91.6</u>
#1 + GRPO	43.7	72.9	50.7	**92.7**
#1 + SFT+GRPO (Ours)	**71.1**	**84.7**	**75.9**	88.7

C. Commentary Refinement Evaluation

Exp. 3: Segment-level commentary baseline comparison. Tab. III shows that GAMESIGHT has competitive performance across both settings. It has relatively lower similarity to live-text commentary compared to MatchVision in BLEU and CIDER, since the knowledgeable commentary has a naturally lower similarity to the original live-text commentary, as shown in the last row under *Live-text as GT*. They are usually longer and more analytical, differs significantly in natural language style. While, GAMESIGHT outperforms others when using live televised as GT, showing a better alignment to the human-centric TV commentary.

TABLE III: Comparison with baselines.

Model	**BLEU**	**METEOR**	**ROUGE**	**CIDEr**
Live text as GT				
MatchVoice	15.182	22.142	18.262	<u>13.514</u>
MatchVision	**28.311**	<u>25.827</u>	<u>25.173</u>	**27.143**
GAMESIGHT **(ours)**	<u>20.320</u>	**27.160**	**27.968**	0.887
Live televised	4.038	11.778	6.272	0.001
Live televised as GT				
MatchVoice	<u>8.615</u>	**10.067**	8.247	0.113
MatchVision	4.214	7.945	<u>8.946</u>	<u>0.369</u>
GAMESIGHT **(ours)**	**17.409**	<u>9.051</u>	**10.497**	**3.442**

Exp. 4: Segment-level accuracy of knowledge reference. As shown in Fig. IV, SoccerKAG improves external knowledge accuracy by 16.11% compared to SoccerRAG, which primarily attributes to the enhanced query exemplars and time constrains in query schema. Regarding the internal game context, we further break down the test into two categories: goal-related information (ICL_{goal}) and other context aspects (ICL_{other}). GAMESIGHT presents high accuracy particularly in goal events, as the goal information is specifically quantified by the score. A slight drop in accuracy is observed when referencing other internal context elements, for instance, fouls and corner kicks, as the model independently accounts for various demanded statistics.

TABLE IV: Results on knowledge reference accuracy.

	$SoccerRAG$	$SoccerKAG$	ICL_{goal}	ICL_{other}
Acc. $\uparrow$	64.60	**81.80**	**98.76**	**90.74**

Exp. 5: Game-level quantitative results in commentary analysis. *Coh-Metrix* [24] is a discourse analyze tool that includes various indicators to evaluate discourse coherence. With the focus on contextual relevance, we adopted two key indicators, deep coherence and anaphor overlap. *Deep coherence* reflects the degree of text containing intentional connectors when causal and logical relationships present. As shown in Tab. V, televised and GAMESIGHT commentary have higher deep coherence that helps readers form a more engaging and in-depth understanding of causal events, processes, and behaviors during the game. *Anaphor overlap* measures the overlap between nouns and pronouns in adjacent sentences, indicating the semantic continuity within the passage by pointing back to the context. Televised and GAMESIGHT commentary show higher anaphor overlap, indicating better contextual relevance. *Sentiment Score* [33] is a polarity score reveals

the sentiment polarity (1=positive, -1=negative). Live text has almost neutral sentiment, while televised and GAMESIGHT commentary provide positive and vivid atmosphere. The MOS test also shows that audience prefers our commentary than the original live text version.

TABLE V: Results in commentary analysis. DC, AO, SS refer to *Deep Cohesion*, *Anaphor Overlap*, and *Sentiment Score*.

Method	DC↑	AO↑	SS↑	MOS↑
MatchVoice	0.556	0.164	0.159	2.27
MatchVision	0.631	0.156	0.066	2.72
Live Text	0.474	0.174	-0.059	3.14
GAMESIGHT **(ours)**	**0.842**	0.213	0.194	4.08
Live Televised	0.813	**0.234**	**0.254**	**4.22**

Exp. 6: Game-level commentary structural similarity. As shown in Fig. 6, while live text commentary remains small proportion of explanation and commentary, GAMESIGHT has a similar structure with live televised commentary, which meets the criteria from [7] with less than 50% description, indicating the depth and insights in the commentary with appropriate logical composition.

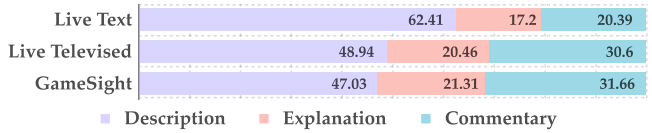

Fig. 6: Commentary structural composition in percentage.

VI. CONCLUSION

In this work, we proposed GAMESIGHT to tackle soccer commentary generation as a knowledge-enhanced visual reasoning task. It enables automatic commentary generation with precise entity and contextual knowledge, which leans towards the real-world live televised commentary that provides the audience with informative and engaging experience. We believe our work paved way for leveraging video-LMMs in fine-grained sports analysis.

VII. ACKNOWLEDGMENTS

This work is supported by the National Natural Science Foundation of China Nos. 62425604 and 62502256. It is also supported by Beijing Natural Science Foundation (L257006), High Performance Computing Center, Tsinghua University and Beijing Zitiao Network Technology Co., Ltd.

REFERENCES

[1] FIFA, "The football landscape," 2021.
[2] Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck, "SoccerNet-caption: Dense video captioning for soccer broadcasts commentaries," 2023.
[3] Cise Midoglu, Steven A. Hicks, Vajira Thambawita, Tomas Kupka, and Pål Halvorsen, "MMSys'22 grand challenge on AI-based video production for soccer," 2022.
[4] Jiayuan Rao, Haoning Wu, Chang Liu, and Weidi Xie, "MatchTime: Towards automatic soccer game commentary generation," 2024.
[5] Xiang Li, Yangfan He, Shuaishuai Zu, Zhengyang Li, Tianyu Shi, Yiting Xie, and Kevin Zhang, "Multi-modal large language model with RAG strategies in soccer commentary generation," 2025.
[6] Ji Qi, Jifan Yu, Teng Tu, Kunyu Gao, Yifan Xu, and Xinyu Guan, "GOAL: A challenging knowledge-grounded video captioning benchmark for real-time soccer commentary generation," 2023.
[7] Desheng Zhang, Feng Li, and Ziye Wang, "Three kinds of logic on sport interpretation & commentary and their scientific application," *Journal of Shanghai University of Sport*, vol. 41, no. 2, pp. 15–20, 2017, [J].
[8] Tom Hedrick, *The art of sportscasting: How to build a successful career*, Taylor Trade Publications, 2000.
[9] Vladimir Somers, Victor Joos, Silvio Giancola, Anthony Cioppa, and Seyed Abolfazl Ghasemzadeh, "SoccerNet game state reconstruction: End-to-end athlete tracking and identification on a minimap," June 2024.
[10] Muhammad Rafiq, Ghazala Rafiq, Rockson Agyeman, Gyu Sang Choi, and Seong-Il Jin, "Scene classification for sports video summarization using transfer learning," *Sensors*, vol. 20, no. 6, pp. 1702, 2020.
[11] Jiayuan Rao, Haoning Wu, Hao Jiang, Ya Zhang, Yanfeng Wang, and Weidi Xie, "Towards universal soccer video understanding," 2024.
[12] Eric Zelikman, Yuhuai Wu, Jesse Mu, and Noah D. Goodman, "STaR: Bootstrapping reasoning with reasoning," 2022.
[13] Zhengbao Jiang, Frank F. Xu, Luyu Gao, Zhiqing Sun, Qian Liu, Jane Dwivedi-Yu, Yiming Yang, Jamie Callan, and Graham Neubig, "Active retrieval augmented generation," 2023.
[14] Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A Smith, and Mike Lewis, "Measuring and narrowing the compositionality gap in language models," *arXiv preprint arXiv:2210.03350*, 2022.
[15] Aleksander Theo Strand, Sushant Gautam, Cise Midoglu, and Pål Halvorsen, "SoccerRAG: Multimodal soccer information retrieval via natural queries," 2024.
[16] Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen, "Lora: Low-rank adaptation of large language models," 2021.
[17] Xiaodong Wang and Peixi Peng, "Open-r1-video," https://github.com/Wang-Xiaodong1899/Open-R1-Video, 2025.
[18] 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*, 2002, pp. 311–318.
[19] Ramakrishna Vedantam, "Cider: Consensus-based image description evaluation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 4566–4575.
[20] Satanjeev Banerjee, "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*, 2005, pp. 65–72.
[21] Chin-Yew Lin, "ROUGE: A package for automatic evaluation of summaries," Tech. Rep. W04-1013, ACL Workshop on Text Summarization Branches Out, 2004, https://aclanthology.org/W04-1013.
[22] Adrien Deliège, Anthony Cioppa, and Silvio Giancola, "SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos," 2021.
[23] Sushant Gautam, "SoccerNet-echoes: A soccer game audio commentary dataset," 2024.
[24] Danielle S McNamara, Arthur C Graesser, Philip M McCarthy, and Zhiqiang Cai, *Automated evaluation of text and discourse with CohMetrix*, Cambridge University Press, 2014.
[25] Qwen Team, "Qwen2.5-vl," January 2025.
[26] Jinguo Zhu, Weiyun Wang, Zhe Chen, Zhaoyang Liu, and Shenglong Ye, "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," 2025.
[27] Boqiang Zhang, Kehan Li, Zesen Cheng, Zhiqiang Hu, and Yuqian Yuan, "Videollama 3: Frontier multimodal foundation models for image and video understanding," 2025.
[28] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, and Chunyuan Li, "Llava-onevision: Easy visual task transfer," 2024.
[29] Gemini Team, Google, "Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities," 2025.
[30] OpenAI, "Hello gpt-4o," 2024, Accessed: 2025-04-10.
[31] OpenAI, "Introducing openai o1-preview," 2024, Accessed: 2025-04-10.
[32] DeepSeek-AI and Daya Guo et al., "Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning," 2025.
[33] C.J. Hutto and E. Gilbert, "Vader: A parsimonious rule-based model for sentiment analysis of social media text," in *Proceedings of the Eighth International Conference on Weblogs and Social Media*, 2014.