SoccerRef-Agents: Multi-Agent System for Automated Soccer Refereeing

Zi Meng[1,2], Wanli Song[1], Yi Hu[2], Jiayuan Rao[†2], and Gang Chen[†2]

[1] University of Michigan, Michigan, USA
{mengzi,wanlis}@umich.edu
[2] Shanghai Jiao Tong University, Shanghai, China
{huyi_0811, jy_rao, chengang76}@sjtu.edu.cn

Abstract. Refereeing is vital in sports, where fair, accurate, and explainable decisions are fundamental. While intelligent assistant technologies are being widely adopted in soccer refereeing, current AI-assisted approaches remain preliminary. Existing research mostly focuses on isolated video perception tasks and lacks the ability to understand and reason about foul scenarios. To fill this gap, we propose **SoccerRef-Agents**, a holistic and explainable multi-agent decision-making framework for soccer refereeing. The main contributions are: (i) constructing the multimodal benchmark **SoccerRefBench** with over 1,200 referee theory questions and 600 foul video clips; (ii) building a vector-based knowledge base **RefKnowledgeDB** using the latest *"Laws of the Game"* and a classic case database for precise, knowledge-driven reasoning; (iii) designing a novel multi-agent architecture that collaborates via cross-modal RAG to bridge the semantic gap between visual content and regulatory texts. This work explores the technical capability of integrating MLLMs with refereeing expertise, and evaluations show our system significantly outperforms general-purpose MLLMs in decision accuracy and explanation quality. All databases, benchmarks, and code will be made available.

Keywords: AI Referee · Sports Understanding · Multi-Agent System · Multimodal Reasoning.

1 Introduction

Competitive sports captivate global audiences with their inherent dynamism and passion. Among them, soccer, hailed as the most beautiful game, enjoys unparalleled worldwide attention. In modern soccer, with advancements in science and technology, particularly the recent development of AI, technology is being applied across all aspects from training and competition to broadcasting. Within this context, soccer refereeing occupies a special position, as it is tasked with the crucial work of maintaining the orderly conduct of matches. The fairness and integrity of a match are highly dependent on the accuracy of refereeing decisions. Currently, the latest advancements in artificial intelligence have been applied

[†] Corresponding author.

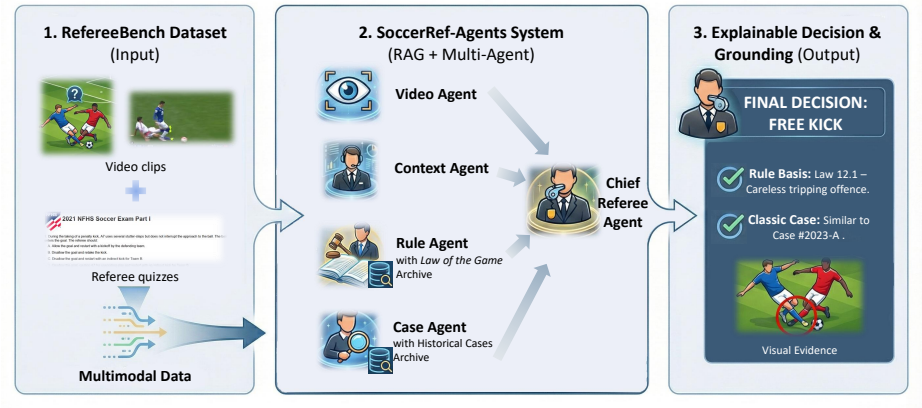

Fig. 1. Overview of **SoccerRef-Agents**. The system mimics a professional officiating team by decomposing the task into perception (Video Agent), background analysis (Context Agent), legal interpretation (Rule Agent), and precedent retrieval (Case Agent), culminating in a final decision by the Chief Referee Agent.

to many aspects of soccer understanding, including video comprehension [38, 37, 20, 39, 43, 28] However, in the domain of automated refereeing, despite some attempts, three main challenges remain: **(i) The Gap Between Perception and Reasoning:** Most existing systems[14, 15] treat foul detection as a simple visual classification task. However, effective officiating requires a cognitive leap from visual data to rule-based logic. For instance, determining disciplinary actions such as distinguishing between "reckless" and "using excessive force" is a knowledge-intensive process that necessitates explicit reference to the *"Laws of the Game(LOTG)"*[33], a capability largely absent in current vision-only models. **(ii) Unreliable and Unexplainable Outputs:** General-purpose multimodal large models often struggle to provide reliable justifications for their decisions. Due to a lack of internalized domain knowledge, these models[3, 5, 16] may generate seemingly plausible but factually incorrect explanations, known as hallucinations. They typically fail to ground their judgments in specific rules or similar historical cases, limiting their practical utility as transparent decision-support systems. **(iii) Limitations in Sample Size:** Most current researches[12, 41] focus on analyzing and processing small-scale samples, lacking a truly large-scale benchmark for the comprehensive evaluation of AI refereeing systems.

To bridge this gap, we propose a comprehensive framework for holistic and explainable soccer refereeing. Recognizing that a valid judgment requires both accurate perception and authoritative knowledge, we first construct **SoccerRef-Bench**, a multimodal benchmark specifically designed to evaluate refereeing logic. Unlike general sports datasets, **SoccerRefBench** integrates theoretical knowledge derived from professional certification exams with practical judgment from controversial broadcast replay clips. Furthermore, to enable knowledge-driven reasoning, we construct **RefKnowledgeDB** by digitizing and vectoriz-

ing the latest *Laws of the Game*[33] and a curated database of classic historical cases, serving as the long-term memory of our system.

To tackle the complexity of decision-making, we introduce **SoccerRef-Agents**, a novel multi-agent system that mimics the collaborative workflow of a professional officiating team. As illustrated in Figure 1, our system decomposes the refereeing task into specialized roles: a *Video Agent* for perception, a *Rule Agent* for legal interpretation, a *Case Agent* for precedent retrieval, and a *Chief Referee Agent* for final adjudication. By leveraging a cross-modal Retrieval-Augmented Generation (RAG) mechanism, our agents can consult the rulebook based on visual analysis, ensuring decisions are not only accurate but also legally cited.

Concretely, we make the following contributions in this paper:

(i) We construct **SoccerRefBench**, a specialized multimodal benchmark for soccer officiating. It comprises 1,218 theoretical exam questions and 600 annotated video clips of fouls, mapped to standard disciplinary outcomes (No Offence, Normal, Yellow, Red).
(ii) We establish a specialized vector knowledge base: the **RefKnowledgeDB**. The database enables precise and fine-grained retrieval of regulations and historical precedents, mitigating model hallucinations.
(iii) We introduce **SoccerRef-Agents**, a multi-agent framework featuring a novel cross-modal reasoning pipeline. By utilizing video descriptions to drive textual knowledge retrieval, our system effectively bridges the semantic gap between visual footage and legal texts.
(iv) Extensive evaluations on **SoccerRefBench** demonstrate the superiority of our agentic system. **SoccerRef-Agents** significantly outperforms general-purpose MLLMs in decision accuracy and achieves state-of-the-art performance in generating legally grounded explanations.

2 Related Works

2.1 Sports Understanding

Sports understanding has emerged as a pivotal testbed for evaluating Multimodal Large Language Models (MLLMs) due to its dynamic nature and complex rules. Early research primarily focused on atomic tasks such as action recognition[37, 20, 39, 43] and automated scoring[40, 29, 36]. With the advent of large-scale benchmarks[8, 13, 42], the focus has shifted towards holistic understanding. For instance, Sports-QA[19] and SPORTU[35] evaluate models on QA tasks across multiple sports disciplines, requiring agents to perceive actions and reason about game states. Recent works have also explored dense video captioning[22, 25] and commentary generation[38, 28] to enhance fan engagement. However, these general benchmarks often treat sports rules as implicit background knowledge rather than explicit logic constraints, limiting their ability to evaluate fine-grained professional judgment.

2.2 Soccer Understanding

As the world's most popular sport, soccer has attracted significant research attention. The field is largely driven by extensive datasets like SoccerNet[9] , which facilitates tasks ranging from action spotting to replay grounding. Beyond perception, recent studies have ventured into higher-level cognitive tasks. MatchTime[28] and SoccerNet-Caption[22] focus on generating time-aligned commentary, while other works explore tactical analysis[24, 18] and game state reconstruction[30]. despite these advancements, existing soccer understanding models predominantly focus on descriptive tasks[27, 26] or statistical analysis[2, 17, 6]. They rarely address the normative aspect of the game—specifically, judging player actions against the strict, textual framework of the Laws of the Game[33].

2.3 Soccer Refereeing

The automation of refereeing has advanced significantly in both industrial applications and academic research, albeit with distinct focuses. In professional leagues, hardware-centric systems like Video Assistant Referee (VAR) [31] and Semi-Automated Offside Technology (SAOT) [7] provide precise spatial measurements. However, they lack the semantic reasoning capabilities required to judge subjective fouls or interpret player intent. Conversely, academic research focuses on vision-based foul recognition from broadcast footage. For instance, VARS [14] treats refereeing as a multi-view classification problem, while X-VARS [15] explores utilizing MLLMs to generate descriptive explanations for foul events.

Despite these strides, a fully autonomous AI Referee remains elusive. Recent trends, such as the SoccerNet Challenges [4, 10], predominantly frame officiating as a video classification task, mapping visual features directly to disciplinary labels based on statistical correlations. However, authentic refereeing is a complex legal adjudication process, not merely a perceptual one. Existing models [4, 10, 34] fail to logically derive verdicts by verifying factual preconditions and aligning them with specific regulatory clauses. This disconnect between visual perception and legal reasoning limits the reliability and explainability of current approaches. To bridge this gap, effective AI systems must move beyond pattern recognition to knowledge-driven reasoning. However, existing datasets lack the explicit alignment between model intuition and the strict textual framework of the *Laws of the Game*[33] required to develop and evaluate this capability. This necessity motivates the construction of **SoccerRefBench**, establishing the foundational resources detailed in Sec. 3

3 Dataset Construction

To bridge the critical gap between model intuition and legal adjudication in automated officiating, we introduce a comprehensive data framework designed for knowledge-driven reasoning. This section first outlines the motivation and

overview of our proposed framework in Sec. 3.1. We then provide detailed descriptions of the data collection and curation processes for our multimodal benchmark, **SoccerRefBench**, in Sec. 3.2 and Sec. 3.3, respectively. Finally, we elaborate on the construction of our specialized domain repository, the **RefKnowledgeDB**, in Sec. 3.4, which serves as the foundation for the system's retrieval capabilities.

3.1 Motivation & Overview

Soccer refereeing is a highly specialized domain that requires precise interpretation of the rule and split-second decision-making. While existing benchmarks primarily focus on action recognition or general sports question-answering, they still lack evaluation standards specifically tailored for soccer refereeing, particularly in multimodal understanding and rule-based reasoning. To provide an evaluation platform for such a professional scenario, we introduce **SoccerRefBench**, a comprehensive multimodal benchmark designed to assess automated officiating according to professional refereeing standards, comprises two distinct modalities: (i) a textual subset assessing theoretical knowledge via standardized exams, and (ii) a video subset for evaluating practical judgment in controversial foul scenarios. To enable better reasoning and evidence retrieval in the automated officiating pipeline, we construct a searchable knowledge base **RefKnowledgeDB**, containing the official *Laws of the Game*[33] and historical precedents, thereby supporting knowledge-driven reasoning.

3.2 Data Collection

To construct a comprehensive evaluation benchmark, we aggregate data from authoritative refereeing examinations and large-scale video datasets. The overall data collection and aggregation pipeline is illustrated in Figure 2.

3.2.1 Textual Data Collection. We collect a total of 1,218 multiple-choice questions covering diverse aspects of refereeing theory.

- **Chinese Referee Exams:** We select 118 high-quality questions from the Chinese Football Association (CFA) referee certification exams.
- **International Standards:** We collect 1,100 questions from publicly available referee certification exams conducted by the National Federation of State High School Associations (NFHS) and the Florida High School Athletic Association (FHSAA). To ensure temporal coverage and rule evolution adaptability, the dataset spans multiple years, including 2013, 2017, 2019, 2020, 2021, 2022, 2023, and 2025 for NFHS and 2016 for FHSAA.

3.2.2 Video Data Collection. For the visual component, we leverage the SoccerNet-MVFoul dataset[14], which provides multi-view video clips of foul incidents. We extract 600 representative samples that feature clear refereeing controversies.

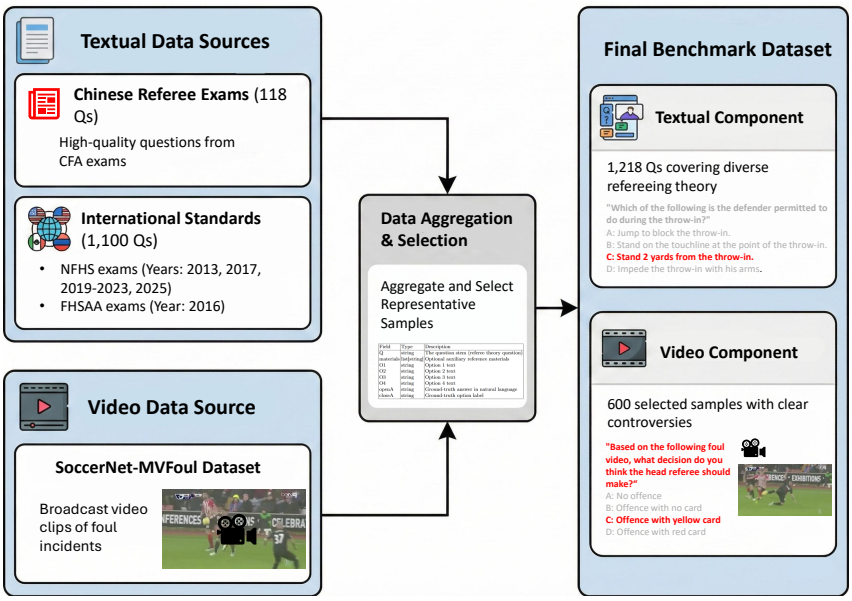

Fig. 2. Overview of the data collection pipeline for **SoccerRefBench**. The dataset integrates 1,218 textual theory questions from diverse international sources and 600 video-based judgment scenarios derived from the SoccerNet-MVFoul dataset.

3.3 Data Curation

To facilitate standardized and objective evaluation, we curate the collected raw data into a unified multiple-choice question format. Each entry in **SoccerRefBench** consists of a question stem, two to four candidate options, and exactly one correct ground-truth answer. This consistent closed-set QA structure across both textual and video modalities allows for precise accuracy metrics and direct comparison between different models.

For the textual component, data processing is mainly translation. For questions sourced from Chinese Referee Exams, we translated the content into English using LLM-assisted tools, followed by manual verification by domain experts to ensure terminological precision. For the National Federation of State High School Associations (NFHS) and Florida High School Athletic Association (FHSAA) exams, we retained the original official options.

For the video component, we transform the raw annotations from SoccerNet-MVFoul[14] into decision-making tasks. To align with the professional referee's disciplinary logic, we map the original ground-truth labels into four distinct severity levels: *No Offence*, *Normal Offence*, *Offence with Yellow Card*, and *Offence with Red Card*. For each video sample, the mapped ground truth serves as the correct option, while the remaining three severity levels are automatically

assigned as distractors. This design compels the model to not merely recognize an action, but to adjudicate its severity against strict officiating standards.

3.4 Knowledge Base Construction

To support accurate decision-making in subsequent intelligent officiating, we construct a specialized domain repository named **RefKnowledgeDB**. This repository comprises two distinct vector databases designed to assist the automated pipeline in retrieval and reasoning.

3.4.1 Laws of the Game (LOTG) Knowledge Base ($\mathcal{K}_{rules}$). We construct the rule-based knowledge base using the latest edition of the IFAB Laws of the Game (2025/26)[33]. To preserve structural integrity and ensure precise citation, we parse the raw PDF documents and perform segmentation at the page level. Each page segment is enriched with metadata and encoded into a high-dimensional vector space using OpenAI's embedding model. This page-level granularity ensures that the retrieved context maintains the semantic coherence of specific regulations.

3.4.2 Classic Case Knowledge Base ($\mathcal{K}_{cases}$). To support high-level Case-Based Reasoning (CBR), we curate a specialized knowledge base comprising 184 historical incidents. These cases are primarily sourced from elite European leagues and the FIFA World Cup, capturing a wide spectrum of standard and controversial officiating scenarios. Each entry is structured in a JSON format that includes a detailed `Case Description`, the official `Decision`, and the perceived `Controversiality` level. This fused text is vectorized to form the retrieval index, while structured attributes are stored as metadata. This design enables the system to retrieve precedents based on semantic similarity across both scenario details and decision outcomes. For a comprehensive breakdown of the data format and source statistics, refer to Appendix B.

4 Methodology

We present **SoccerRef-Agents**, a multi-modal, multi-agent framework designed to mimic the cognitive decision-making process of professional football referees. Unlike traditional end-to-end models, our system decomposes the officiating task into perception, retrieval, legal interpretation, and final adjudication. In this section, we first formulate the problem (Sec. 4.1), show our multi-agent architecture(Sec. 4.2) and detail the agentic workflow for both textual and video-based scenarios (Sec. 4.3).

4.1 Problem Formulation

We define the refereeing task as a conditional generation problem that requires both a discrete decision and a legally grounded explanation. Given a query input q consisting of a textual description q_{txt} or in some cases, together with

a video clip q_{vid}, the system aims to generate a final decision y and a corresponding explanation e. The decision y denotes the predicted outcome, and it represents the correct option for theoretical questions or the disciplinary action for video scenarios. Complementing this, e provides a comprehensive justification that synthesizes the specific input evidence with the Laws of the Game[33] and historical precedents. Based on the above definitions, the refereeing task can be formally expressed as follows:

$$(y, e) = \mathcal{F}(q; \mathcal{K}_{rules}, \mathcal{K}_{cases}),$$

where $\mathcal{F}$ represents the **SoccerRef-Agents** system, and $\mathcal{K}$ denotes the external knowledge bases as mentioned in Sec. 3.4.

4.2 Multi-Agent Architecture

The core of our framework is a collaborative ecosystem comprising four specialized agents and one central decision-maker. Each agent is designed to handle a distinct aspect of the refereeing workflow, from perception to legal reasoning. To mitigate hallucinations and ensure legally reasoning, the agents are supported by two external vectorized knowledge bases that serve as long-term memory.

4.2.1 Agents. Each agent plays a complementary role, collectively emulating the collaborative decision-making process of a professional refereeing team:

- **Video Agent:** A Vision-Language Model (VLM) specialist responsible for parsing raw video frames into structured textual descriptions and providing initial option recommendations based on visual perception.
- **Rule Agent:** Responsible for interpreting the *Laws of the Game*[33]. It takes the retrieved rules and summarizes their applicability to the current scenario, ensuring decisions are grounded in the official statute.
- **Case Agent:** Specializes in historical analogy. It compares the current situation with retrieved classic cases to provide precedent-based insights, helping to resolve ambiguities through similar past judgments.
- **Context Agent:** Extracts and narrates the match importance (e.g., Derby, Final, League match) provided in the metadata text. This contextual information is crucial for judging subjective factors like "intent" and "game management."
- **Chief Referee Agent:** The final decision-maker. It aggregates the outputs from all subordinate agents (Video, Rule, Case, and Context) to synthesize a comprehensive rationale and derive the final verdict.

4.2.2 Knowledge Retrieval Mechanism. Among the aforementioned agents, the *Rule Agent* and *Case Agent* serve as the most direct and essential components for football rule-based decision-making. To support these two specific agents with precise domain knowledge, we implement a Retrieval-Augmented Generation

(RAG) mechanism. This process retrieves the most relevant knowledge segments based on the semantic similarity between the query and the database entries.

To measure the semantic relevance, we utilize cosine similarity. The retrieval process extracts the top-K most relevant segments by maximizing this similarity score. Formally, given a query q' (which can be a text question or a generated video description) and a knowledge base $\mathcal{K}$ (either $\mathcal{K}_{rules}$ or $\mathcal{K}_{cases}$), the retrieval set $\mathcal{R}$ is defined as:

$$\mathcal{R}(q', \mathcal{K}) = \underset{k \in \mathcal{K}}{\text{TopK}} \left(\text{Sim}(\mathbf{E}(q'), \mathbf{E}(k)) \right) \tag{1}$$

where $\mathbf{E}(\cdot)$ is the embedding function mapping inputs to a high-dimensional vector space. The similarity function $\text{Sim}(\cdot, \cdot)$ is explicitly calculated as the dot product of the normalized vectors:

$$\text{Sim}(\mathbf{u}, \mathbf{v}) = \frac{\mathbf{u} \cdot \mathbf{v}}{\|\mathbf{u}\| \|\mathbf{v}\|} = \frac{\sum_{i=1}^{d} u_i v_i}{\sqrt{\sum_{i=1}^{d} u_i^2} \sqrt{\sum_{i=1}^{d} v_i^2}} \tag{2}$$

Here, $\mathbf{u} = \mathbf{E}(q')$ and $\mathbf{v} = \mathbf{E}(k)$ represent the embedding vectors of the query and the knowledge segment, respectively. The TopK operator selects the K segments with the highest similarity scores, which are then passed to the downstream agents for summarization and reasoning.

4.3 Workflow and Reasoning Chains

Given the variations in the form and focus of the queries we receive, we build upon previous 5 agents to construct a collaborative multi-agent system, which is designed with two distinct reasoning pipelines based on the input modality.

4.3.1 Text-Mode Reasoning Pipeline. When the Router identifies the input as text, the workflow proceeds as follows: (i)*Retrieval:* The input text q_{txt} is used directly as the query to retrieve the top-3 relevant segments from both $\mathcal{K}_{rules}$ and $\mathcal{K}_{cases}$. (ii)*Specialist Analysis:* The *Rule Agent* generates a rule summary (S_{rule}) and a concise logic chain linking the text to the rule based on the retrieved laws. Simultaneously, the *Case Agent* generates a precedent summary (S_{case}) based on similar historical scenarios. (iii)*Final Adjudication:* The *Chief Referee Agent* receives the original text q_{txt}, S_{rule}, and S_{case}. It synthesizes this legal and historical context to select the correct option and generate an explanation.

$$S_{rule} = \mathcal{A}_{rule}(\mathcal{R}(q_{txt}, \mathcal{K}_{rules}))$$

$$S_{case} = \mathcal{A}_{case}(\mathcal{R}(q_{txt}, \mathcal{K}_{cases}))$$

$$(y, e) = \mathcal{A}_{chief}(q_{txt}, S_{rule}, S_{case})$$

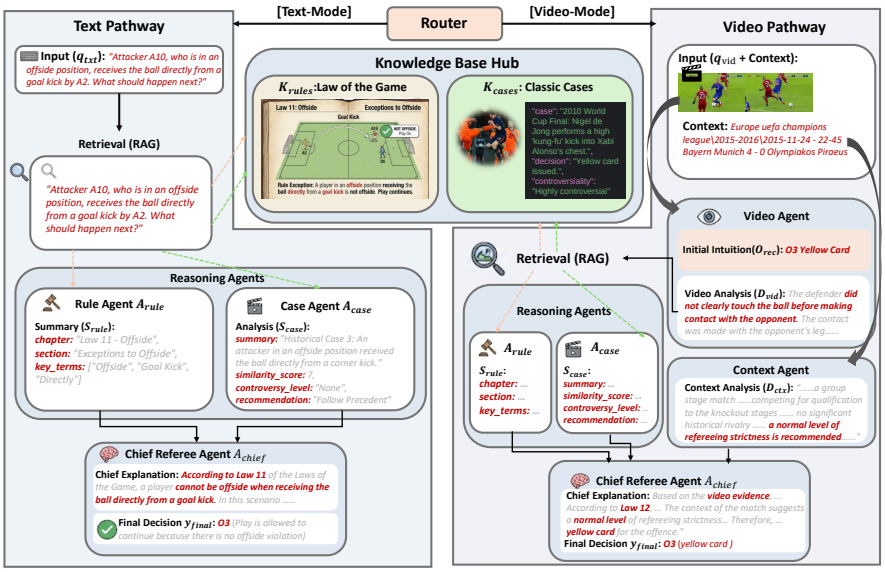

Fig. 3. Overview of the system's dual-pathway reasoning workflow. Depending on the input modality identified by the Router, the system executes either the Text-Mode or Video-Mode pipeline. The **Text-Mode Pipeline** directly utilizes the input query for retrieval. The **Video-Mode Pipeline** features a specialized Cross-Modal RAG mechanism, where the Video Agent converts visual information into a textual Video Analysis (D_{vid}) to bridge the semantic gap for retrieving laws and cases. Both pipelines culminate in the Chief Referee Agent, which synthesizes specialist summaries (S_{rule}, S_{case}) and comprehensive context to render final adjudication.

4.3.2 Video-Mode Reasoning Pipeline. When the Router identifies the input as video, the workflow proceeds as follows: (i) *Visual Perception:* The *Video Agent* processes the video q_{vid}. It outputs two key components: a preliminary recommendation (O_{rec}) and a detailed Video Analysis (D_{vid}), which describes the player actions, contact point, and intensity in natural language. (ii) *Contextualization:* The *Context Agent* processes the accompanying context text q_{ctx} to generate a match background description (D_{ctx}). (iii) *Cross-Modal Retrieval:* Crucially, we use the generated Video Analysis D_{vid} as the query embedding for the RAG module. This bridges the modality gap, allowing us to retrieve relevant rules and cases based on the visual narrative.The *Rule Agent* produces S_{rule} based on D_{vid} driven retrieval, and the *Case Agent* produces S_{case} based on D_{vid} driven retrieval. (iv) *Aggregation & Decision:* The *Chief Referee Agent* synthesizes the comprehensive information set: $\{O_{rec}, D_{vid}, D_{ctx}, S_{rule}, S_{case}\}$. This allows the Chief Referee to validate the *Video Agent*'s initial intuition against strict rules and precedents before making the final decision.

$$D_{vid}, O_{rec} = \mathcal{A}_{video}(q_{vid})$$
$$D_{ctx} = \mathcal{A}_{context}(q_{ctx})$$
$$S_{rule} = \mathcal{A}_{rule}(\mathcal{R}(D_{vid}, \mathcal{K}_{rules}))$$
$$S_{case} = \mathcal{A}_{case}(\mathcal{R}(D_{vid}, \mathcal{K}_{cases}))$$
$$(y, e) = \mathcal{A}_{chief}(O_{rec}, D_{vid}, D_{ctx}, S_{rule}, S_{case})$$

5 Experiments

Based on the multi-agent system we have constructed, we conducted extensive experiments to evaluate its performance. This section outlines the experimental protocols in Sec. 5.1, reports comprehensive quantitative comparisons against state-of-the-art MLLMs in Sec. 5.2 and Sec. 5.3, shows human evaluation in Sec. 5.4, and concludes with qualitative analysis in Sec. 5.5.

5.1 Experimental Settings

5.1.1 Baselines. We evaluate our **SoccerRef-Agents** against a comprehensive suite of state-of-the-art MLLMs on **SoccerRefBench**, covering both proprietary commercial APIs and open-source models. *(i) Commercial APIs:* We leading commercial APIs with top-tier performance on broad multimodal and reasoning tasks, including GPT-4o[23], Claude 4.5 Sonnet[1] and Gemini 2.5 Flash[11]. These API models are prompted with detailed system instructions but rely solely on their internal parametric knowledge (i.e., without external retrieval). *(ii) Open-Source Models:* We evaluate the Qwen3-VL series (8B and 32B)[32] as competitive public VLMs for both text and video understanding. Additionally, for the text-only task, we include DeepSeek-V3[21] as a strong textual baseline.

5.1.2 Metrics. As for the evaluation on our proposed benchmark **SoccerRef-Bench**, we employ Accuracy (Acc) as the primary quantitative metric for both subsets. To further assess the interpretability and practical validity of the system's decisions, we introduce a human evaluation protocol. In this protocol, evaluations are conducted by individuals with professional refereeing experience. The details of this protocol are elaborated in Section 5.4.

5.2 Quantitative Results

We present the comparative results on **SoccerRefBench** in Table 1.Our observations are as follows: (i) The results on **SoccerRefBench** exhibit a broad performance spectrum, effectively differentiating the capabilities of various MLLMs. Accuracy scores span from 46.88% to 79.56% in the text domain and 23.50% to 40.17% in the video domain. This significant variance underscores the complexity and diversity of our benchmark, confirming that it serves as a rigorous

Table 1. Main Results on **SoccerRefBench**. We compare **SoccerRef-Agents** with state-of-the-art commercial and open-source MLLMs. The best performance is highlighted in **bold**, and the second best is underlined. Note that DeepSeek-V3[21] is a text-only model and is thus evaluated only on the Text subset.

Model	**Text**(%)	**Video**(%)	**Overall**(%)
Open-Source Models			
Qwen3-VL-8B[32]	46.88	23.50	39.16
Qwen3-VL-32B[32]	56.57	24.33	45.93
DeepSeek-V3[21]	65.35	-	-
Commercial APIs			
Gemini 2.5 Flash[11]	69.38	26.33	55.17
Claude 4.5 Sonnet[1]	65.19	34.67	55.12
GPT-4o[23]	77.83	37.67	64.58
SoccerRef-Agents (Ours)	**79.56**	**40.17**	**66.56**

testing ground capable of distinguishing the performance gaps between standard open-source models, advanced commercial APIs, and specialized agentic systems. (ii) As shown in Table 1, **SoccerRef-Agents** achieves the highest accuracy across both modalities. In the text domain, our system reaches **79.56%**, surpassing the strongest baseline (GPT-4o[23]) by 1.73%. This indicates that even highly capable generalist models like GPT-4o struggle with the professional soccer referee theories and highly specific knowledge required for professional referee exams, whereas our RAG-enhanced *Rule Agent* effectively closes this gap. (iii) The video domain task proves to be extremely challenging for all models, with no baseline exceeding 38% accuracy. This highlights the intrinsic difficulty of distinguishing subtle foul grades solely from broadcast footage. Despite this, **SoccerRef-Agents** achieves state-of-the-art performance with **40.17%** accuracy, outperforming GPT-4o[23] (+2.5%) and significantly surpassing open-source video models like Qwen3-VL[32] (+15.8%). This performance gain validates our cross-modal design, where the system aligns visual descriptions with textual rules and historical knowledge to make more informed decisions.

5.3 Ablation Studies

To further investigate the individual contributions of our specialized knowledge bases and the synergy between the *Rule Agent* and *Case Agent*, we conduct ablation experiments on the **SoccerRefBench**. We evaluate three configurations: (i) **SoccerRef-Agents (Full)**, our complete multi-agent framework; (ii) **w/o Rule Agent**, where the system relies solely on historical precedents and internal knowledge; and (iii) **w/o Case Agent**, where the system only retrieves information from the *Laws of the Game*[33].

Table 2. Ablation Study of Knowledge Components. We report the Accuracy (%) on the Text (1,218 Qs), Video (600 Qs) subsets, and the weighted Overall performance.

Variant	Text	Video	Overall
w/o Rule Agent	78.90%	**42.50%**	**66.89%**
w/o Case Agent	**79.89%**	39.17%	66.45%
SoccerRef-Agents (Full)	79.56%	40.17%	66.56%

The results in Table 2 reveal a notable divergence in knowledge utilization across modalities. The text task exhibits rule-heavy characteristics, achieving peak accuracy (**79.89%**) when relying solely on the *Rule Agent*; here, historical cases act as noise that interferes with strictly literal rule interpretations. Conversely, the video task is inherently case-heavy, with the *w/o Rule Agent* variant reaching the highest accuracy (**42.50%**). This suggests that practical officiating requires the guide of precedents to navigate subjective visual scenarios where rigid rule-following may prove too restrictive. Although individual components yield marginal gains in specialized domains, our full **SoccerRef-Agents** framework provides the most robust balance across both modalities. Crucially, as shown in our human evaluation, integrating both the rule guideline and historical precedents is essential for generating the professionally justifiable and legally grounded explanations required for multi-modal refereeing.

5.4 Human Evaluation of Explanation Quality

Accuracy alone does not fully capture the utility of an AI referee; the reasoning process must be transparent, authoritative, and legally binding. Since current automatic metrics correlate poorly with logical validity in this specialized domain, we conduct a rigorous human evaluation campaign.

We focus our evaluation on comparing **SoccerRef-Agents** against the top-performing baseline, **GPT-4o**[23]. We randomly sample 100 textual questions and 50 video questions from our **SoccerRefBench**. We invite three certified referees authorized by the Chinese Football Association to act as expert annotators. To ensure objectivity, the evaluation is conducted in a blind setting where the source of each explanation is masked.Unlike general readability scores, our scoring rubric is strictly tailored to professional officiating standards. Experts rate each explanation on a 1-5 Likert scale whose definition can be checked in the appendix D.

The comparative evaluation results from three expert referees are detailed in Table 3. While GPT-4o[23] exhibits strong linguistic fluency, experts noted it frequently falters in rule adherence, inventing plausible-sounding but non-existent regulations. In contrast, our system effectively minimizes such domain-specific hallucinations. By explicitly citing the *Laws of the Game*[33] and anchoring reasoning in historical precedents, **SoccerRef-Agents** provides more trustworthy

Table 3. Human Evaluation Results. Three professional referees rated the explanation quality (1-5 scale). Columns denote average scores for Text (100 Qs), Video (50 Qs), and Overall weighted performance.

Model	Referee 1			Referee 2			Referee 3			Average		
	Text	Video	Overall	Text	Video	Overall	Text	Video	Overall	Text	Video	Overall
GPT-4o[23]	3.63	2.70	3.32	3.65	2.70	3.33	3.72	2.80	3.41	3.67	2.73	3.36
Ours	**3.76**	**2.76**	**3.43**	**4.09**	**3.24**	**3.81**	**3.96**	**3.18**	**3.70**	**3.94**	**3.06**	**3.65**

and legally grounded explanations, closely mirroring professional cognitive workflows.

5.5 Qualitative Analysis

To demonstrate the interpretability and reasoning depth of **SoccerRef-Agents**, we present a visualization of the decision-making process in Figure 4. Unlike black-box models that output a label without context, our system provides a transparent "glass-box" view of its logic through detailed `agent_traces`.

As illustrated in Figure 4, our framework effectively handles diverse modalities. For theoretical questions, the system accurately retrieves specific chapters from the *Laws of the Game*. For instance, in the second column regarding a disqualified coach, the *Rule Agent* correctly pinpoints "Law 3 - The Players" and "Extra persons on the field of play". Simultaneously, the *Case Agent* retrieves a historical precedent with a high similarity score (9/10), reinforcing the *Chief Referee Agent*'s decision to "terminate the game" rather than merely delaying it. In complex visual scenarios, the *Video Agent* acts as the primary perceiver. In the third column, the agent explicitly notes that "the video frames do not clearly show any [contact]". This visual evidence drives the *Chief Referee Agent* to conclude "No Offence," aligning with the ground truth. Also, in the video questions, the *Context Agent* identifies the match such as "league match in the Bundesliga between Wolfsburg and Dortmund". These information provides match context for the *Chief Referee Agent* to confirm the standard of judgment. This granular visibility into the reasoning process confirms that **SoccerRef-Agents** does not merely memorize answers but actively constructs a legal argument, significantly enhancing trust in automated officiating. More results are provided in the Appendix E.

6 Conclusion

This paper presents a comprehensive framework for explainable and standardized automated soccer refereeing for multimodal scenarios. Concretely, we introduce **SoccerRefBench**, a unique multimodal benchmark bridging theoretical exams and practical video judgments, accompanied by a specialized vector knowledge base **RefKnowledgeDB** to ensure legal precision. To tackle the complexity of officiating logic, we establish **SoccerRef-Agents**, a novel multi-agent system

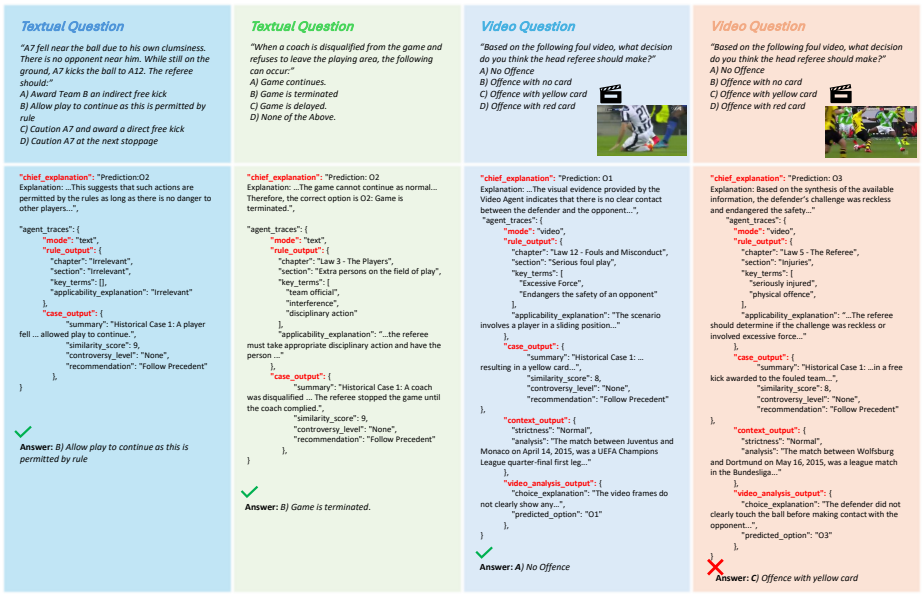

Fig. 4. Qualitative examples of **SoccerRef-Agents** on **SoccerRefBench**. The figure illustrates the step-by-step reasoning chains for both textual (left two columns) and video-based (right two columns) queries. Key intermediate outputs from the Rule Agent, Case Agent, and Video Agent are explicitly logged in `agent_traces`, providing legally grounded justifications for the final decision.

that mimics the collaborative workflow of professional referee teams. By leveraging a cross-modal Retrieval-Augmented Generation (RAG) mechanism, our system effectively translates visual evidence into rule-based reasoning, bridging the gap between perception and adjudication. Extensive evaluations have demonstrated the superiority of our framework, showing that it not only achieves higher accuracy than general-purpose models but also provides transparent, citation-backed explanations. We believe this work establishes a new foundation for fair, transparent, and interpretable AI assistants in sports officiating.

References

1. Anthropic: Claude 4.5 sonnet. https://www.anthropic.com (2025)
2. Bai, L., Gedik, R., Egilmez, G.: What does it take to win or lose a soccer game? a machine learning approach to understand the impact of game and team statistics. Journal of the Operational Research Society **74**(7), 1690–1711 (2023)
3. Bai, Z., Wang, P., Xiao, T., He, T., Han, Z., Zhang, Z., Shou, M.Z.: Hallucination of multimodal large language models: A survey. arXiv preprint arXiv:2404.18930 (2024)
4. Cioppa, A., Giancola, S., Somers, V., Joos, V., Magera, F., Held, J., Ghasemzadeh, S.A., Zhou, X., Seweryn, K., Kowalczyk, M., Mr'oz, Z., Lukasik, S., Halo'n, M., Mkhallati, H., Deliège, A., Hinojosa, C., Sanchez, K., Mansourian, A.M., Miralles, P., Barnich, O., Vleeschouwer, C.D., Alahi, A., Ghanem, B., Droogenbroeck, M.V., Gorski, A., Clapés, A., Boiarov, A., Afanasiev, A., Xarles, A., Scott, A., Lim, B., Yeung, C., Gonzalez, C., Rüfenacht, D., Pacilio, E., Deuser, F., Altawijri, F.S., Cach'on, F., Kim, H.B., Wang, H., Choe, H., Kim, H.J., Kim, I.M., Kang, J.M., Tursunboev, J., Yang, J., Hong, J., Lee, J., Zhang, J., Lee, J., Zhang, K., Habel, K., Jiao, L., Li, L., Guti'errez-P'erez, M., Ortega, M., Li, M., Lopatto, M., Kasatkin, N., Nemtsev, N., Oswald, N., Udin, O., Kononov, P., Geng, P., Alotaibi, S.G., Kim, S., Ulasen, S., Escalera, S., Zhang, S., Yang, S., Moon, S., Moeslund, T.B., Shandyba, V., Golovkin, V., Dai, W., Chung, W., Liu, X., Zhu, Y., Kim, Y., Li, Y., Yang, Y., Xiao, Y., Cheng, Z., Li, Z.: Soccernet 2024 challenges results. ArXiv **abs/2409.10587** (2024), https://api.semanticscholar.org/CorpusID:272693834
5. Dong, B., Ni, M., Huang, Z., Yang, G., Zuo, W., Zhang, L.: Mirage: Assessing hallucination in multimodal reasoning chains of mllm. arXiv preprint arXiv:2505.24238 (2025)
6. Elmiligi, H., Saad, S.: Predicting the outcome of soccer matches using machine learning and statistical analysis. In: 2022 IEEE 12th Annual Computing and Communication Workshop and Conference (CCWC). pp. 1–8. IEEE (2022)
7. FIFA: Semi-automated offside technology. Inside FIFA (July 17 2023), https://inside.fifa.com/innovation/world-cup-2022/semi-automated-offside-technology, accessed: 2026-01-13
8. Fu, C., Dai, Y., Luo, Y., Li, L., Ren, S., Zhang, R., Wang, Z., Zhou, C., Shen, Y., Zhang, M., et al.: Video-mme: The first-ever comprehensive evaluation benchmark of multi-modal llms in video analysis. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 24108–24118 (2025)
9. Giancola, S., Amine, M., Dghaily, T., Ghanem, B.: Soccernet: A scalable dataset for action spotting in soccer videos. In: Proceedings of the IEEE conference on computer vision and pattern recognition workshops. pp. 1711–1721 (2018)
10. Giancola, S., Cioppa, A., Guti'errez-P'erez, M., Held, J., Hinojosa, C., Joos, V., Leduc, A., Magera, F., Sanchez, K., Somers, V., Xarles, A., Agudo, A., Alahi, A., Barnich, O., Clap'es, A., Vleeschouwer, C.D., Escalera, S., Ghanem, B., Moeslund, T.B., Droogenbroeck, M.V., Abe, T., Alotaibi, S.G., Altawijri, F.S., Araujo, S., Bai, X., Bi, X., Cao, J., Chao, V., Czarnog'orski, K., Deuser, F., Du, M., Feng, T., Frenzel, P., Fuchs, M., Garc'ia, J., Habel, K., Hashiguchi, T., Hirose, S., Hu, X., Hwang, Y., Inoue, R., Itsuji, R., Iwai, K., Ji, H., Ji, Y., Jiao, L., Kageyama, Y., Kamikawa, Y., Kanasugi, Y., Kim, H., Kim, J., Kurihara, T., Li, B., Li, L., Li, X., Lian, Y., Liang, D., Lin, H., Lin, J., Liu, J., Liu, L., Liu, S., Liu, Z., Lu, Y., M'endez, F., Ma, H., Ma, W., Maksymiuk, J., Mantilla, H., Mathkour, I., Matthes, D., Motomochi, A., Muhammad, A.R., Nakayama, H., Oh, J., Oo, Y.M., Ortega, M., Oswald, N.,

Otsubo, R., Perez, F., Qi, M., Rey, C., Reyes-Angulo, A., Rose, O., Rueda-Chac'on, H., Saito, H., Sarmiento, J., Sawafuji, K., Scott, A., Shen, X., Shrestha, P., Sim, J.Y., Sun, L., Sun, Y., Suzuki, T., Tang, L., Tonouchi, M., Uchida, I., Velesaca, H.O., Wang, T., Watanabe, R., Wu, J., Wu, Y., Yamagishi, S., Yang, D., Yang, X., Yang, Y., Ye, H., Ye, X., Yeung, C., Yu, X., Zhang, C., Zhang, D., Zhang, K., Zhao, Z., Zhou, X., Zhu, W., Ziegler, J.: Soccernet 2025 challenges results. ArXiv **abs/2508.19182** (2025), https://api.semanticscholar.org/CorpusID:280870241

11. Google: Gemini 2.5 flash. https://deepmind.google (2025)
12. Gottschalk, C., Tewes, S., Niestroj, B., Jäger, C., Drees, J., Ernst, A.: Innovation in elite refereeing through ai technological support for dogso decisions. International Journal of Operations Management **2**(3), 7–15 (2022)
13. He, X., Feng, W., Zheng, K., Lu, Y., Zhu, W., Li, J., Fan, Y., Wang, J., Li, L., Yang, Z., et al.: Mmworld: Towards multi-discipline multi-faceted world model evaluation in videos. arXiv preprint arXiv:2406.08407 (2024)
14. Held, J., Cioppa, A., Giancola, S., Hamdi, A., Ghanem, B., Van Droogenbroeck, M.: Vars: Video assistant referee system for automated soccer decision making from multiple views. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5086–5097 (2023)
15. Held, J., Itani, H., Cioppa, A., Giancola, S., Ghanem, B., Van Droogenbroeck, M.: X-vars: Introducing explainability in football refereeing with multi-modal large language models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3267–3279 (2024)
16. Kuang, J., Shen, Y., Xie, J., Luo, H., Xu, Z., Li, R., Li, Y., Cheng, X., Lin, X., Han, Y.: Natural language understanding and inference with mllm in visual question answering: A survey. ACM Computing Surveys **57**(8), 1–36 (2025)
17. Kubayi, A., Larkin, P.: Match-related statistics differentiating winning and losing teams at the 2019 africa cup of nations soccer championship. Frontiers in Sports and Active Living **4**, 807198 (2022)
18. Lee, G.J., Jung, J.J.: Dnn-based multi-output model for predicting soccer team tactics. PeerJ Computer Science **8**, e853 (2022)
19. Li, H., Deng, A., Ke, Q., Liu, J., Rahmani, H., Guo, Y., Schiele, B., Chen, C.: Sports-qa: A large-scale video question answering benchmark for complex and professional sports. arXiv preprint arXiv:2401.01505 (2024)
20. Li, Q., Chiu, T.C., Huang, H.W., Sun, M.T., Ku, W.S.: Videobadminton: a video dataset for badminton action recognition. In: 2024 IEEE International Conference on Big Data (BigData). pp. 1387–1392. IEEE (2024)
21. Liu, A., Feng, B., Xue, B., Wang, B., Wu, B., Lu, C., Zhao, C., Deng, C., Zhang, C., Ruan, C., et al.: Deepseek-v3 technical report. arXiv preprint arXiv:2412.19437 (2024)
22. Mkhallati, H., Cioppa, A., Giancola, S., Ghanem, B., Van Droogenbroeck, M.: Soccernet-caption: Dense video captioning for soccer broadcasts commentaries. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 5074–5085 (2023)
23. OpenAI: Gpt-4o. https://openai.com (2024)
24. PLAKIAS, S., KOKKOTIS, C., GIAKAS, G., TSAOPOULOS, D., MOUSTAKIDIS, S.: Can artificial intelligence revolutionize soccer tactical analysis? Trends in Sport Sciences **31**(3) (2024)
25. Qi, M., Wang, Y., Li, A., Luo, J.: Sports video captioning via attentive motion representation and group relationship modeling. IEEE Transactions on Circuits and Systems for Video Technology **30**(8), 2617–2633 (2019)

26. Rao, J., Li, Z., Wu, H., Zhang, Y., Wang, Y., Xie, W.: Multi-agent system for comprehensive soccer understanding. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3654–3663 (2025)
27. Rao, J., Wu, H., Jiang, H., Zhang, Y., Wang, Y., Xie, W.: Towards universal soccer video understanding. In: Proceedings of the Computer Vision and Pattern Recognition Conference. pp. 8384–8394 (2025)
28. Rao, J., Wu, H., Liu, C., Wang, Y., Xie, W.: Matchtime: Towards automatic soccer game commentary generation. arXiv preprint arXiv:2406.18530 (2024)
29. Shao, D., Zhao, Y., Dai, B., Lin, D.: Finegym: A hierarchical video dataset for fine-grained action understanding. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2616–2625 (2020)
30. Somers, V., Joos, V., Cioppa, A., Giancola, S., Ghasemzadeh, S.A., Magera, F., Standaert, B., Mansourian, A.M., Zhou, X., Kasaei, S., et al.: Soccernet game state reconstruction: End-to-end athlete tracking and identification on a minimap. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3293–3305 (2024)
31. Spitz, J., Wagemans, J., Memmert, D., Williams, A.M., Helsen, W.F.: Video assistant referees (var): The impact of technology on decision making in association football referees. Journal of Sports Sciences **39**(2), 147–153 (2021)
32. Team, Q.: Qwen3-vl. https://github.com/QwenLM/Qwen3-VL (2025)
33. The International Football Association Board: Laws of the Game 2025/26. The International Football Association Board, Zurich, Switzerland (2025), https://www.theifab.com, effective from 1st July 2025
34. Wang, J., Li, L.: A method for feature division of soccer foul actions based on salience image semantics. PloS one **20**(6), e0322889 (2025)
35. Xia, H., Yang, Z., Zou, J., Tracy, R., Wang, Y., Lu, C., Lai, C., He, Y., Shao, X., Xie, Z., et al.: Sportu: A comprehensive sports understanding benchmark for multimodal large language models. arXiv preprint arXiv:2410.08474 (2024)
36. Xu, J., Rao, Y., Yu, X., Chen, G., Zhou, J., Lu, J.: Finediving: A fine-grained dataset for procedure-aware action quality assessment. In: Proceedings of the IEEE/CVF conference on computer vision and pattern recognition. pp. 2949–2958 (2022)
37. Xu, J., Zhao, G., Yin, S., Zhou, W., Peng, Y.: Finesports: A multi-person hierarchical sports video dataset for fine-grained action understanding. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 21773–21782 (2024)
38. You, L., Huang, W., Xie, X., Wei, X., Li, B., Lin, S., Li, Y., Wang, C.: Timesoccer: An end-to-end multimodal large language model for soccer commentary generation. In: Proceedings of the 33rd ACM International Conference on Multimedia. pp. 3418–3427 (2025)
39. Yuan, H., Ni, D., Wang, M.: Spatio-temporal dynamic inference network for group activity recognition. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 7476–7485 (2021)
40. Zahan, S., Hassan, G.M., Mian, A.: Learning sparse temporal video mapping for action quality assessment in floor gymnastics. IEEE Transactions on Instrumentation and Measurement **73**, 1–11 (2024)
41. Zhekambayeva, M., Yerekesheva, M., Ramashov, N., Seidakhmetov, Y., Kulambayev, B.: Designing an artificial intelligence-powered video assistant referee system for team sports using computer vision. Retos: nuevas tendencias en educación física, deporte y recreación (61), 1162–1170 (2024)

42. Zhou, J., Shu, Y., Zhao, B., Wu, B., Xiao, S., Yang, X., Xiong, Y., Zhang, B., Huang, T., Liu, Z.: Mlvu: A comprehensive benchmark for multi-task long video understanding. arXiv e-prints pp. arXiv–2406 (2024)
43. Zhu, K., Wong, A., McPhee, J.: Fencenet: Fine-grained footwork recognition in fencing. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 3589–3598 (2022)

A SoccerRefBench Dataset Details

To ensure seamless multimodal evaluation, both textual and video-based queries in **SoccerRefBench** follow a standardized JSON schema shown in Table 4. The primary difference lies in the `materials` field, which for video samples includes the local file path and match context metadata. Listing 1.1 provides examples of the video judgment task and textual task.

Table 4. Internal JSON Schema of the **SoccerRefBench** Dataset.

Field	Type	Description
Q	string	The question stem (referee theory question)
materials	list[string]	Optional auxiliary reference materials
O1	string	Option 1 text
O2	string	Option 2 text
O3	string	Option 3 text
O4	string	Option 4 text
openA	string	Ground-truth answer in natural language
closeA	string	Ground-truth option label

Listing 1.1. SoccerRefBench Dataset Example

```
{
    "Q": "Based on the following foul video, what decision do
        you think the head referee should make?",
    "materials": [
      {
        "path": "Dataset/video/SoccerNet/mvfouls/train/
            action_620/clip_1.mp4",
        "context": "europe_uefa-champions-league
            \\2014-2015\\2015-04-14 - 21-45 Juventus 1 - 0
            Monaco"
      }
    ],
    "openA": "Offence with no card",
    "closeA": "O2",
    "O1": "No offence",
    "O2": "Offence with no card",
    "O3": "Offence with yellow card",
    "O4": "Offence with possible red card"
},
{
    "Q": "Player A1 kicks off to start the second half of the
        game. Player A1's kick goes directly into Team B's
       goal. The referee should:",
    "materials": [
      "none"
```

```
    ],
    "openA": "Award the goal and restart the match with a
        kickoff for Team B.",
    "closeA": "O4",
    "O1": "Disallow the goal and have Team A retake the
        kickoff.",
    "O2": "Disallow the goal and have Team A take an indirect
         free kick from the halfway line.",
    "O3": "Disallow the goal and award Team B a goal kick.",
    "O4": "Award the goal and restart the match with a
        kickoff for Team B."
}
```

B Classic Case Knowledge Base Details

B.1 Raw Case Format

Before vectorization into the **RefKnowledgeDB**, historical cases are curated in a structured format capturing the incident, the official decision, and the perceived controversiality. Listing 1.2 showcases the raw entries.

Listing 1.2. Structured format of historical cases in the Classic Case Knowledge Base.

```
{
  "id": 4,
  "case": "2024 Premier League: Declan Rice receives a second
      yellow card for slightly kicking the ball away to
     delay the restart.",
  "decision": "Second yellow card and red card issued.",
  "controversiality": "Highly controversial"
},
{
    "id": 61,
    "case": "2012 UCL: John Terry knees Alexis Sanchez in the
         back during an off-the-ball incident.",
    "decision": "Red card issued.",
    "controversiality": "Non-controversial"
},
{
    "id": 179,
    "case": "2021 La Liga: Referee calls players back from
        the locker room to take a penalty after VAR review.",
    "decision": "Penalty awarded.",
    "controversiality": "Somewhat controversial"
}
```

B.2 Source Statistics

The cases are aggregated from authoritative tournament archives and officiating review panels. Table 5 summarizes the distribution of historical precedents.

Table 5. Distribution of sources for the Classic Case Database.

Source Authority	Number of Cases
FIFA World Cup	72
Premier League	40
UEFA Champions League	24
Bundesliga	19
La Liga	17
Euro Cup	12

C Multi-Agent System Prompts

In this section, we provide the specialized system prompts for each agent in **SoccerRef-Agents**.

Rule Agent System Prompt

You are an expert AI Legal Analyst specializing in the IFAB Laws of the Game. Your task is to strictly analyze the provided rule excerpts and identify the exact clause that governs the user's scenario.

Rule Agent User Prompt

Context (Retrieved IFAB Laws): `{retrieved_rules}`
Scenario: *"{query_ text}"*

Instructions:
- Analyze the specific Law, Section, and Bullet Point.
- Prioritize specific offenses (e.g., **Serious Foul Play**).
- Extract text verbatim if relevant.

Expected Output: JSON format with `direct_quote`, `key_terminology_match`, and `confidence` fields.

Context Agent User Prompt

Match Context: `{context_str}`

Instructions:
- Analyze the match importance (e.g., Derby, Final, League match), home/-away factors, and potential team rivalries.
- Determine the recommended refereeing strictness (Lenient, Normal, Strict).

Expected Output (JSON):
A JSON object containing:
- `strictness`: Recommended enforcement level.
- `analysis`: Brief justification based on match atmosphere and stakes.

Video Agent System Prompt

You are a professional AI Soccer Referee Assistant. The input contains video frames from a **live soccer match broadcast replay**. Output JSON only.

Video Agent User Prompt

Input Type: Broadcast Replay Video

Question: `{question_text}`

Options:
`{options_str}`

Instructions:
- Analyze the input video carefully.
- Select the ONE correct option ID.
- Provide a brief explanation in English.

Expected Output (JSON ONLY):

```
{
  "choice_explanation": "...",
  "predicted_option": "O1"
}
```

Chief Referee Agent System Prompt

You are the Chief Referee Agent, the final decision-maker in a multi-agent soccer refereeing system. Your role is to synthesize evidence from specialized subordinate agents (Rule, Case, Context, and Video) to provide a definitive ruling on complex foul scenarios.

Chief Referee Agent User Prompt

=== QUESTION DATA ===
Question: `{question_text}`
Options: `{options_text}`
PS: 'A#' means player of team A with jersey number #, same for 'B#'.

=== SUBORDINATE AGENT REPORTS ===
[1] Reference Law:
`{rule_str_placeholder}` *%(Includes Text of Law, Match Logic, and Confidence)%*

[2] Reference Precedents:
`{case_str_placeholder}` *%(Includes Valid Precedent or No Precedent status)%*

[3] Reference Context (Video Mode Only):
`{context_analysis}`

[4] Visual Evidence (Video Agent):

- **Video Agent's Choice Explanation:** `{desc}`
- **Video Agent's Initial Intuition:** `{pred}`

=== INSTRUCTIONS ===
- Analyze the provided input text and subordinate reports carefully.
- Select the most correct ONE option ID.
- Provide a brief explanation in English.

OUTPUT FORMAT:
```
Prediction: [Option ID]
Explanation: [Reasoning]
```

D Details on human evaluation of explanation quality

We created an html webpage for the professional referees to score the quality of explanation output by the model

D.1 The definition of Likert Scale

- **5 - Perfect:** Precisely identifies the foul or answers the theory question with correct legal citations. The causal logic is clear, and the terminology is professional.
- **4 - Good:** The verdict and key descriptions are correct. May contain minor terminological imprecisions or verbose explanations, but the core reasoning is valid.
- **3 - Fair:** The conclusion is correct, but the explanation contains slight hallucinations or misses rule support.
- **2 - Poor:** Cites incorrect rules or describes actions significantly inconsistent with the video evidence.

- **1 - Nonsense:** Contains severe hallucinations or chaotic logic that contradicts basic football common sense.

D.2 Human Evaluation Interface

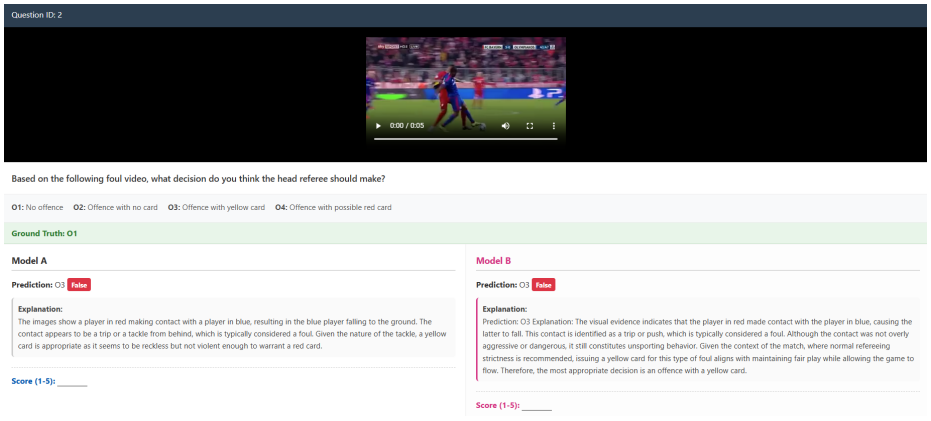

Fig. 5. Human Evaluation Interface

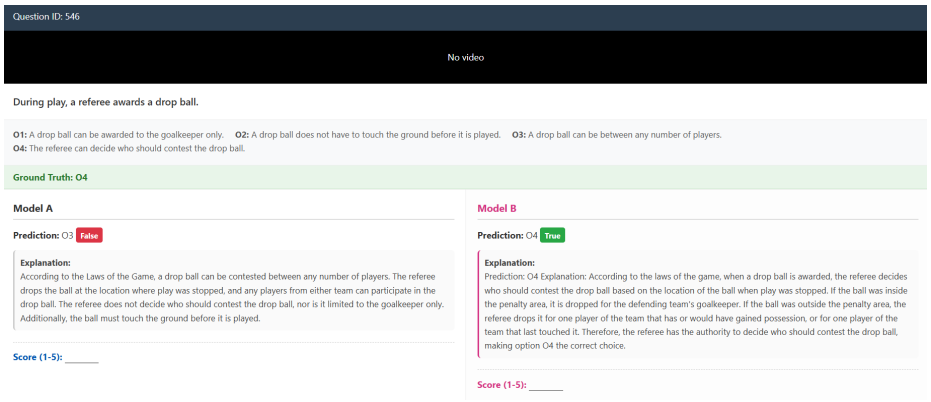

Fig. 6. Human Evaluation Interface

E Additional Qualitative Results

Figure 7 and Figure 8 present further examples of **SoccerRef-Agents**'s performance on **SoccerRefBench**.

Fig. 7. Extended qualitative results

Fig. 8. Extended qualitative results