

# X-VARS: Introducing Explainability in Football Refereeing with Multi-Modal Large Language Models

Jan Held<sup>1</sup> Hani Itani<sup>2</sup> Anthony Cioppa<sup>1</sup> Silvio Giancola<sup>2</sup>  
 Bernard Ghanem<sup>2</sup> Marc Van Droogenbroeck<sup>1</sup>  
<sup>1</sup> University of Liège    <sup>2</sup> KAUST

## Abstract

The rapid advancement of artificial intelligence has led to significant improvements in automated decision-making. However, the increased performance of models often comes at the cost of explainability and transparency of their decision-making processes. In this paper, we investigate the capabilities of large language models to explain decisions, using football refereeing as a testing ground, given its decision complexity and subjectivity. We introduce the *EXplainable Video Assistant Referee System*, X-VARS, a multi-modal large language model designed for understanding football videos from the point of view of a referee. X-VARS can perform a multitude of tasks, including video description, question answering, action recognition, and conducting meaningful conversations based on video content and in accordance with the Laws of the Game for football referees. We validate X-VARS on our novel dataset, SoccerNet-XFoul, which consists of more than 22k video-question-answer triplets annotated by over 70 experienced football referees. Our experiments and human study illustrate the impressive capabilities of X-VARS in interpreting complex football clips. Furthermore, we highlight the potential of X-VARS to reach human performance and support football referees in the future. We will provide code, model, dataset, and demo upon publication.

## 1. Introduction

Recently, the field of Artificial Intelligence (AI) has witnessed remarkable progress in the development of large language models [3, 6, 16, 25, 64]. These models have acquired a strong language understanding, enabling them to tackle a broad range of linguistic tasks, ranging from text generation and conversation to zero-shot question answering. The development has further progressed with multi-modal language models, going beyond the constraints of text-based inputs, but also including images, videos, and audio [1, 41, 69, 71]. However, as the capabilities of these

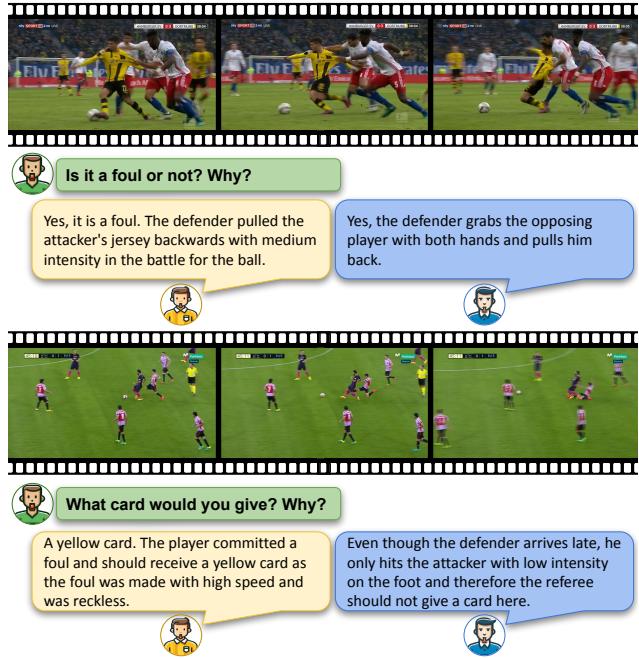


Figure 1. **SoccerNet-XFoul dataset.** Examples of annotations from two different referees for the same foul. The second example illustrates the complexity and subjectivity of refereeing decisions.

models continue to advance, the increase in model performance often comes at the cost of reduced explainability and transparency. This trend poses several challenges for users and developers who seek to understand the *why* and *how* behind a model’s decision-making process. Explaining the reasoning process of AI models is particularly crucial in domains requiring high levels of trust, such as healthcare, autonomous driving, or sports. In the context of football, referee decisions can significantly impact the financial future and existence of clubs, making it essential for AI models to transparently explain their decision-making process. Such transparency is key to building trust and facilitating the acceptance and integration of AI in sports. For instance, during the Qatar World Cup 2022, FIFA employed an AI sys-

tem for semi-automated offside detection [19]. To enhance the system’s transparency and explainability, it generates a 3D representation of the game to allow referees and spectators to visually verify offside positions with undeniable clarity, bridging the gap between AI decision-making and human understanding.

In this paper, we explore the use of large language models to enhance transparency and explainability of automated referee decision-making. Football refereeing offers an ideal environment, as many of the decisions that referees make are subjective and reliant on each individual’s interpretation of the rules of the game. Particularly, we introduce X-VARS, the first multi-modal large language model designed to explain football refereeing decisions. X-VARS is based on a Vision Language Model (VLM) [41] that adapts the design of the LLaVa [37] model for spatio-temporal video modeling. We train X-VARS using a new training paradigm where we input the visual features and the multi-task predictions of our fine-tuned visual encoder CLIP ViT-L/14 [50] to the language model. We validate X-VARS on our novel *SoccerNet-XFoul* dataset containing more than 22k video-question-answer triplets about the most fundamental refereeing questions. More than 70 professional referees annotated our dataset and provided, for each question, detailed explanations about their decisions. X-VARS achieves state-of-the-art performance on the *SoccerNet-MVFoul* dataset, and our human study demonstrates that X-VARS generates explanations for its decisions at a level comparable to human referees. Finally, X-VARS can analyze and understand complex football duels and provide accurate decision explanations, opening doors for future applications to support referees in their decision-making processes.

**Contributions.** We summarize our contributions as follows: **(i)** We publicly release *SoccerNet-XFoul*, a new multi-modal dataset containing more than 22k video-question-answer triplets about refereeing questions. **(ii)** We introduce *X-VARS*, a new vision language model that can perform multiple multi-modal tasks such as visual captioning, question-answering, video action recognition, and can generate explanations of its decisions on-par with human level. **(iii)** We perform a thorough evaluation of our model, including analyses of our new training paradigm, the influence of the CLIP text predictions, and a human study that compares X-VARS to human referees.

## 2. Related work

**Sports video understanding.** The field of sports video understanding has gained a lot of interest lately [63]. It encompasses a wide range of tasks such as player segmentation, detection, and tracking [10, 42, 43, 56, 66], keypoint detection [38], summarizing [20, 45], player re-identification [43, 60], action spotting in untrimmed videos [8, 9, 14, 23, 27,

57–59], pass prediction and feasibility [2, 26], foul recognition [18, 24] or dense video captioning for football broadcasts commentaries [47]. Such tasks can be formulated and solved by leveraging recent advances in deep learning for general video understanding. However, progress in sports video understanding heavily relies on the availability of sports-centric large-scale annotated datasets [31, 32, 44, 49, 54, 65, 70]. *SoccerNet* [7, 11–13, 21, 22] stands among the largest and most comprehensive dataset for video understanding in soccer. *SoccerNet-MVFoul* [24] further extended *SoccerNet* by proposing a novel multi-view football dataset designed for foul classification annotated by professional referees. In this work, we further extend *SoccerNet-MVFoul* into a visual-question-answering dataset focused on football refereeing questions, named *SoccerNet-XFoul*.

**Vision language models.** Natural Language Processing (NLP) has witnessed remarkable advancements with the emergence of open-source Large Language Models (LLMs) [3, 6, 16, 25, 48, 52, 64]. These models have demonstrated exceptional capabilities in language understanding and generation tasks. LLMs have also served as the basis for the success of many vision-language models that are based on projecting the visual features of an image [5, 29, 37, 68, 71] or a video [1, 69] encoder onto the input embedding space of an LLM. This idea and its variants allowed leveraging the power of LLMs for multi-modal understanding. In this work, we propose X-VARS, a vision language model for visual captioning, question-answering, action video recognition, and conducting meaningful conversations based on video content.

**Explainability.** Recently, explainability in machine learning has received lots of attention, leading to the development of various techniques to demystify complex models. LIME (Local Interpretable Model-agnostic Explanations) [51] explains the predictions of any machine learning classifier by approximating any classifier locally with an interpretable model. Meanwhile, SHAP (SHapley Additive exPlanations) [39] offers a unified perspective on feature importance by averaging all feature combinations, ensuring consistent attributions. Grad-CAM (Gradient-weighted Class Activation Mapping) [55] employs gradient information from the final convolutional layer to generate a heat map highlighting crucial input image regions. Counterfactual Explanations [67] identify the least number of changes required in the input data to alter the model’s prediction, offering insights into decision boundaries and feature importance. Lastly, Explanation via Language [33] emphasizes natural language dialogues for enhanced interaction between experts and models, underscoring the importance of interactive systems tailored to user requirements. In this paper, we investigate how large language models can explain decisions, using football refereeing as a testing environment given its decision complexity and subjectivity.

Dataset	Type	#Instances	#Questions	#Context	VQA	Captioning	AR
Conceptual 12M [4]	Images	12M	12M	Various	-	✓	-
LAION-5B [53]	Images	3B	3B	Various	-	✓	-
LLava dataset [37]	Images	158k	158k	Various	✓	✓	-
MovieQA [62]	Videos	408	15k	Movies	✓	-	-
TVQA [34]	Videos	21k	150k	TV shows	✓	-	-
Video Instruction Dataset [41]	Videos	100k	100k	Various	✓	✓	-
HowTo100M [46]	Videos	136M	136M	Youtube	-	✓	-
GOAL [61]	Videos	1k	53k	Football	-	✓	-
Sports-QA [35]	Videos	6k	94k	Football	✓	-	✓
SoccerNet-caption [47]	Videos	942	-	Football games	-	✓	-
<b>SoccerNet-XFoul (Ours)</b>	Football	10k	22k	Football fouls	✓	✓	✓

Table 1. **Comparative overview of relevant datasets.** *SoccerNet-XFoul* contains high-quality answers annotated by more than 70 experienced referees. Our dataset is the largest dataset in sports with complex questions and the only one focusing on refereeing questions. VQA stands for Visual Question Answering, AR stands for Action Recognition.

### 3. SoccerNet-XFoul dataset

The performance of supervised models mostly relies on the quality and quantity of available annotated datasets. Multi-modal datasets are generally harder to curate and annotate, which explains their usually smaller size compared to uncurated datasets. Table 1 shows a comparative overview of multi-modal datasets in the literature, specifically highlighting those that contain combinations of text and image, as well as text and video pairs. We introduce *SoccerNet-XFoul*, a dataset specifically designed for foul video recognition and explanation. It consists of high-quality video-text pairs with more than 10k video clips and 22k questions, annotated by more than 70 experienced referees. Compared to the other sports datasets, *SoccerNet-XFoul* has the most video clips and much more complex questions. In the following, we present our approach to creating this high-quality human-annotated dataset.

**Questions.** We identify 4 key questions on the most foundational, complex, and game-impacting decisions a referee must confront during a game. To answer the two first questions, “*Is it a foul or not? Why?*” and “*What card would you give? Why?*”, the model requires an in-depth understanding of the rules of the game [30] as well as an understanding of the context in which an action occurred. Factors such as the intent, the foul location, the game dynamic and the intensity of the contact must all be considered. The two last questions, if “*the defender stops a promising attack or a goal-scoring opportunity?*” and if “*the referee could have given an advantage?*” add a new layer of difficulty and prediction analysis. The answers to the questions are not only visual, since the model has to make predictions about potential future outcomes. For instance, in assessing whether the referee should have given an advantage, the model needs

to evaluate whether the attacking team would benefit more from continuing play rather than being granted a free-kick.

**Annotators.** As no public dataset is available that provides detailed answers and explanations to these refereeing questions, we conducted an annotation campaign with over 70 referees over a three-month period. To ensure high-quality answers, only experienced referees were selected for the annotations. The referees have officiated between 140 and 2,279 official games, with an average of 655 games. They were allowed to assess as many video clips as they wished, with the flexibility to pause at any time to avoid fatigue. Each annotator had the option to provide answers in German, French, English, or Spanish to prevent any linguistic barriers. The answers were translated from the original language to English by ChatGPT-3.5 [3] and then reviewed by another human referee to ensure accurate translation.

**Subjectivity.** Figure 1 displays an example of the subjectivity of the annotations. While one referee annotator would not give a card because he thought the foul was made with low intensity, the other annotator would give a yellow card because he believed the tackling was made with high speed and was reckless. Due to this inherent subjectivity in refereeing, our objective was to gather multiple answers for the same action, rather than collecting a single decision and explanation for each question. Therefore, the multiple decisions and explanations actually help the model to learn a range of valid interpretations and reasoning strategies employed by human referees. All in all, this can improve the robustness of the AI model, enabling it to make informed decisions even in ambiguous or subjective situations. Practically, the annotators were randomly assigned different video clips, ensuring that the same action might be evaluated multiple times. In the end, for each action, we have, on average, 1.5 answers for the same question.

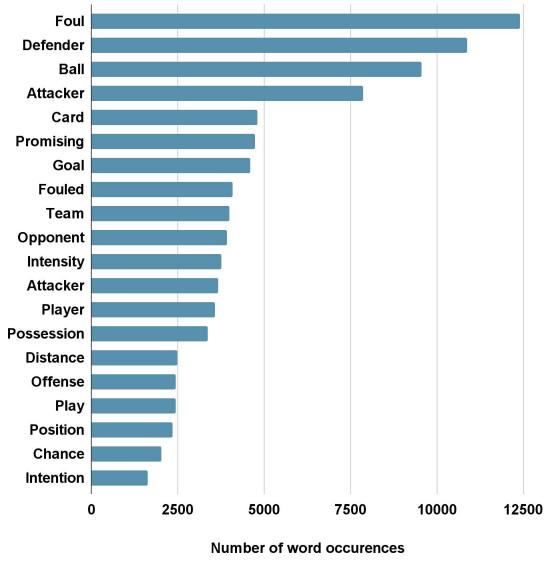


Figure 2. **Distribution of the most common words.** The most frequent words are “foul” and “defender,” followed by semantically related words related to football and referee actions and terms. There is thus a significant imbalance in the distribution.

**Statistics.** Our dataset contains 10k video clips with over 22k referee-generated questions and answers. Figure 2, shows the distribution of the most common words in the explanations of the referee annotators. The most frequently used words are specific terminologies for describing a duel between two players, ranging from descriptive terms such as *defender* or *card* to key terms to consider for evaluating fouls such as *intention* or *intensity*. The number of words per answer ranges from 1 to 66, with a total of more than 540k words and an average of almost 25 words per answer, with a significant imbalance in the distribution of words.

**Novelty.** Compared to traditional visual-question answering datasets, the *SoccerNet-XFoul* dataset is the first to answer refereeing questions, with detailed explanations of why a particular decision is correct. This explainability is a unique approach that enhances the dataset’s complexity and ensures a deeper understanding and representation of real-world scenarios where AI models must make and explain their decisions. Furthermore, the interpretation of situations in our dataset is context-dependent. The interpretation of a foul might differ depending on whether it occurs in the middle of the field or in the penalty area. To correctly answer questions, the model must have an in-depth understanding of the game. Finally, the model needs a level of predictive analysis. For instance, to determine if a defender stopped a promising attack, the model must understand what happened at the moment of the foul and what could have happened in the immediate future. This involves making complex predictions about potential future outcomes, a task that

is far more advanced compared to traditional VQA datasets. Hence, our *SoccerNet-XFoul* dataset is the first and largest visual question-answering dataset for referees in football, offering many new challenges to be explored.

## 4. Methodology

In this section, we provide a comprehensive description of our novel EXplainable Video Assistant Referee System, “X-VARS”, for foul and severity recognition, providing explanations on its decision-making process. We begin by presenting the architecture with a detailed description of each individual component of X-VARS. Then, we provide an in-depth explanation of its training process.

### 4.1. Architecture

Figure 3 illustrates the key architectural components of X-VARS. We use Video-ChatGPT [41], a multi-modal model capable of understanding and generating detailed conversations about videos, as our foundation model. We make several changes to the architecture to adapt it to our needs. Formally, we input a video clip  $\mathbf{v} \in \mathbb{R}^{T \times H \times W \times C}$ , with  $T$ ,  $H$ ,  $W$  and  $C$  being respectively the number of frames, height, width, and channel dimension of the video, to CLIP ViT-L/14 [50],

$$\mathbf{f}_i, \mathbf{h}_i = \text{CLIP}(\mathbf{v}), \quad (1)$$

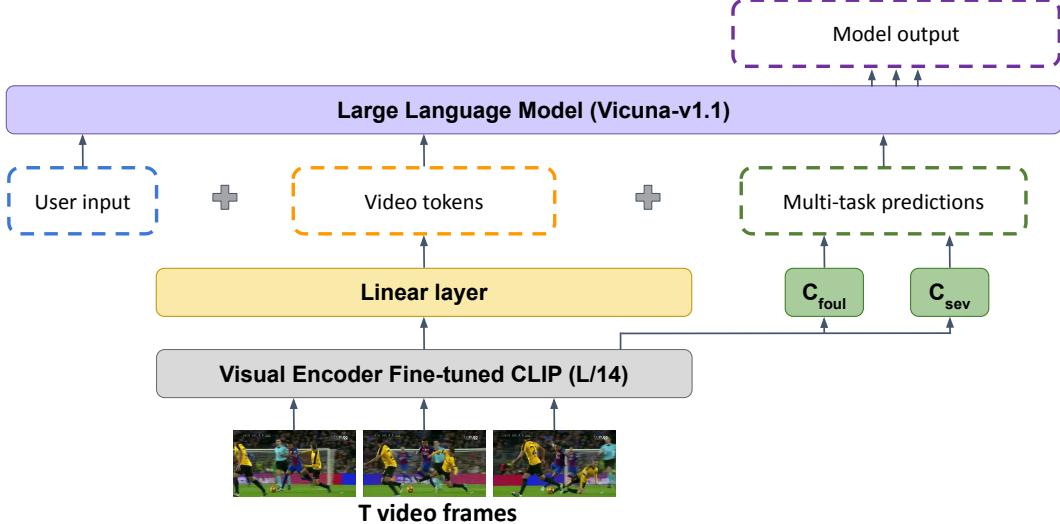
and obtain the corresponding frame feature vector  $\mathbf{f}_i \in \mathbb{R}^{T \times D_1}$  and the hidden states  $\mathbf{h}_i \in \mathbb{R}^{T \times S \times D_2}$ , with  $S$  being the number of tokens obtained by multiplying  $w = W/p$  and  $h = H/p$ , where  $p$  is the patch size of CLIP,  $D_1$  the dimension of the output layer and  $D_2$  the dimension of the hidden states. We then average-pool the hidden states across the temporal dimension to obtain temporal features  $\mathbf{t} \in \mathbb{R}^{S \times D_2}$  and along the spatial dimension to get the video-level spatial representation  $\mathbf{s} \in \mathbb{R}^{T \times D_2}$ . Finally, we concatenate both to obtain spatio-temporal features.

$$\mathbf{z} = [\mathbf{t} \quad \mathbf{s}] \in \mathbb{R}^{(S+T) \times D_2}. \quad (2)$$

Before feeding the video features  $\mathbf{z}$  into the LLM, we project them into the same feature space as the text embeddings by applying a linear projection layer:

$$\mathbf{w} = \text{Linear}(\mathbf{z}) \in \mathbb{R}^{(S+T) \times D_2}. \quad (3)$$

with  $\mathbf{w}$  being a sequence of visual tokens. The feature vectors  $\mathbf{f}_i$  are also average-pooled along the temporal dimension to obtain a single video-level representation  $\mathbf{f} \in \mathbb{R}^{D_1}$ . The video-level feature representation  $\mathbf{f}$  is passed through two classification heads  $\mathbf{C}_{\text{foul}}$  and  $\mathbf{C}_{\text{sev}}$  to obtain the type of foul (*i.e.* Tackling, Holding, Pushing, Standing tackling, Elbowing, Dive, Challenge, or High leg) and to determine whether it is a foul or not, and the corresponding severity



**Figure 3. Architecture of X-VARS.** X-VARS is a visual language model based on a fine-tuned CLIP visual encoder to extract spatio-temporal video features and to obtain multi-task predictions regarding the type and severity of fouls. The linear layer connects the vision encoder to the language model by projecting the video features in the text embedding dimension. We input the projected spatio-temporal features alongside the text predictions obtained by the two classification heads  $C_{foul}$  and  $C_{sev}$  (for the task of determining the type of foul and the task of determining if it is a foul and the corresponding severity) into the Vicuna-v1.1 model, initialized with weights from LLaVA.

(i.e. No offence, Offence + No card, Offence + Yellow card or Offence + Red card), with the predictions being:

$$P_{foul} = \arg \max C_{foul}, \quad (4)$$

$$P_{sev} = \arg \max C_{sev}. \quad (5)$$

These predicted labels with the highest confidence are injected as a textual prompt to the LLM. Hence, this multi-task classification enables the model to utilize acquired information to enhance the performance of the explanation.

To obtain high performance with LLMs, a crucial part consists in determining a prompt which is understandable by the LLM. As we use the Video-ChatGPT backbone, we design our prompt with the following query:

$$USER : < Question >< P_{foul} >< P_{sev} >< w > \quad (6)$$

Assistant :

where  $< Question >$  represents one of our questions randomly sampled from the training set of video-question-answer triplets,  $< P_{foul} >$  and  $< P_{sev} >$  are the two predictions on the foul type and severity recognition task obtained from the fine-tuned CLIP, and  $< w >$  are the projected spatio-temporal features. X-VARS is trained to predict the answers of the assistant as an auto-regressive model.

## 4.2. Training Paradigm

We propose a two-stage training approach. The first stage fine-tunes CLIP on a multi-task classification to learn prior knowledge about football and refereeing. The second

	CLIP	$C_{foul}$ & $C_{sev}$	Linear Layer	LLM
Stage 1	🔥	🔥	-	-
Stage 2	❄️	❄️	🔥	🔥

**Table 2. Overview of the training stages.** In stage 1, we fine-tune CLIP and the classification heads  $C_{foul}$  and  $C_{sev}$  to give X-VARS some prior knowledge about refereeing. In stage 2, we keep them frozen and fine-tune the linear layer and partially the LLM.

step consists in fine-tuning the projection layer and several layers of the LLM to enhance the model’s generation abilities in the sport-specific domain.

**Stage 1: Fine-tuning to inject football knowledge.** While CLIP is excellent at generalizing across various image tasks, it lacks the ability to recognize fine-grained actions or events. These actions are mostly recognizable by considering the temporal dimension rather than images alone. For instance, assessing the severity of a foul requires considering factors such as the intensity and the speed, which cannot be determined by simply examining images. Since CLIP was not trained specifically on football data, the feature representation between two football clips would be very similar, despite the videos depicting different scenarios. Hence, since all our videos are related to football, the output features will share similarities. This proximity between features actually poses a challenge for the LLM, making it difficult to effectively distinguish between different actions. To avoid these issues, we fine-tune CLIP on the *SoccerNet*-

*MVFoul* dataset [24] to learn prior knowledge about football. The training consists of minimizing the summed cross-entropy loss of both tasks. Given the similar magnitudes of both losses, we sum them without scaling or weighting.

**Stage 2: Feature alignment and end-to-end training.** We freeze the fine-tuned CLIPs weights and continue training the linear projection layer and the LLM. Training a projection layer from scratch requires many high-quality video-text pairs and computational resources. To alleviate this, we use the pre-trained weights of the projection layer from Video-ChatGPT [41], which was trained on a dataset of 100k video-text pairs. We further fine-tune this projection layer to map the spatio-temporal features of our football clips into the same dimensional space as the word embeddings. As demonstrated in [37, 40, 41, 71], a simple linear layer is sufficient for connecting video features with the word embedding. During training, we replace the predictions of CLIP  $\langle P_{foul} \rangle$  and  $\langle P_{sev} \rangle$  by the ground truth  $\langle G_{foul} \rangle$  and  $\langle G_{sev} \rangle$  as CLIPs predictions might be noisy, which could lead to confusions. Since determining foul type and severity is subjective, alignment between the ground truth of the *SoccerNet-MVFoul* dataset and the referee responses from our *SoccerNet-XFoul* dataset may vary. Consequently, even when giving  $\langle G_{foul} \rangle$  and  $\langle G_{sev} \rangle$  during training, the model may not only use this information without using the video tokens to produce the text.

## 5. Experiments

In this section, we analyze the performance of X-VARS on the two most important refereeing questions: “*Is it a foul or not? Why?*” and “*What card would you give? Why?*”. Given the importance of these questions, we conduct a comprehensive and detailed analysis, providing insights into the improvements in the video recognition performance, a human study to assess the model’s explanations, some qualitative results, and a thorough ablation study.

### 5.1. Implementation details

We fine-tune CLIP-L/14 on the *SoccerNet-XFoul* dataset for 14 epochs with a learning rate of  $5 \times 10^{-6}$  on a single Nvidia V100 GPU with a batch size of 64, using gradient accumulation to overcome memory limitation. The fine-tuning of the model takes about 9 hours. We use 16 frames in 224p resolution per clip, with 8 frames before and 8 frames after the foul. For the second stage, we employ QLORA [15, 28] to enhance memory efficiency and enable training on a single GPU. We only fine-tune 1% of the layers for 3 epochs using a learning rate of  $2 \times 10^{-4}$  and an overall batch size of 32. The training on 2 A100 40GB GPUs takes about 2 hours. Table 2 provides an overview of the state of the various key components during training.

		Distribution				
	Mean	1	2	3	4	5
Referees	4.0	3%	10%	8%	46%	33%
X-VARS	3.8	3%	17%	4%	46%	30%

Table 3. Score and distributions obtained during our human study comparing the quality of referees and X-VARS generated explanations. The mean scores of X-VARS closely match those of human referees. In 46% of the video clips, X-VARS achieved higher scores for its explanations than the human referees. The distribution of the results is very similar for human referees and X-VARS. A score of 5 is the highest and represents *strongly agree* while 1 is *strongly disagree*.

### 5.2. Human study on explanation performance

Evaluating generative tasks, such as text, image, or video generation, remains a significant challenge due to their subjective nature and the absence of proper evaluation metrics. Traditional language metrics are not very informative for our purpose, as two sentences can be linguistically very similar, yet have entirely different meanings. To achieve quantitative results, we conducted a human study with 20 football referees who evaluated the quality of responses without knowing if they were generated by a human referee or by X-VARS. The referee officiated between 85 and 850 official games, with an average of 490 games. Each participant assessed 20 random video clips, each lasting 5 seconds, with no time restrictions. They evaluated the quality of the explanation, considering whether the evaluation was consistent with the video and if the decision and explanation aligned with the Laws of the Game [30]. They rated each explanation on a scale of 1 to 5, with 5 indicating *strongly agree* and 1 indicating *strongly disagree*. Table 3 shows the results, with X-VARS performing similarly to the human referees, with only minimal score differences. X-VARS’s explanations were more convincing in 46% of the cases than the referee’s. While both show similar results for *strongly agree* and *agree*, X-VARS obtains more *disagree* responses than human referees. The majority of videos where participants disagreed with X-VARS involve types of fouls that are rare in our dataset, *i.e.* when the defender uses his arms illegally by pushing his opponent or hitting him with the elbow in the face. Overall, the human study highlights X-VARS’s impressive ability to understand football videos and explain its decisions at a level comparable to human referees.

### 5.3. Qualitative results

Figure 4 showcases two examples of conversations generated by our proposed X-VARS. Particularly, we illustrate its remarkable ability to understand and generate decisions with explanations related to visual content and the Laws of the Game [30]. Although X-VARS was only fine-tuned on



(a) **Q:** What card would you give? Why? **GT:** No card because the defender briefly held onto the attacker's arm during the fight for the ball, without it being unsportsmanlike.

(b) **Q:** What card would you give? Why? **GT:** No card. Even though the defender had no chance to play the ball, he touched the attacker with low intensity on the foot.

Figure 4. **Qualitative results.** Although X-VARS has never been specifically fine-tuned for conversation, it has inherited its conversational capabilities from the pre-trained model. X-VARS demonstrates impressive discussion skills while being aligned with the video content and the Laws of the Game. (a) X-VARS is close to the ground truth and is able to accurately answer the user’s question. (b) This example shows the subjectivity of foul situations. X-VARS interprets the foul as medium intensity, while the human referee interprets it as low intensity with no chance to play the ball.

two questions, Figure 4 illustrates that X-VARS can generalize and accurately answer or describe video content without any specific fine-tuning. Furthermore, X-VARS was not fine-tuned for conversation, but we inherited these capabilities from the pre-trained conversational model Video-ChatGPT [41], which serves as the foundation for X-VARS. Hence, throughout our two fine-tuning stages, we retained the conversation capabilities of the foundation model and can generate meaningful conversations with X-VARS about football and refereeing. Another interesting fact is the typical characteristic of LLMs to consistently agree with human users. Surprisingly, X-VARS mostly maintains its decision and offers comprehensive explanations for it, even when asked questions such as “*Should the defender receive a red card?*”, when the specific foul would not require any.

However, similarly to other LLMs, X-VARS has also inherited typical issues such as hallucinations, in which it recognizes actions in the video that are not present. Future work could investigate if more high-quality data or more advanced LLMs would limit this hallucination effect.

#### 5.4. Ablation study

**Video action recognition performance.** Table 4 shows the performance of CLIP-L/14 after the fine-tuning process in foul classification tasks. We compare it to the previous state-of-the-art (SOTA) achieved by Held *et al.* [24] by using the same number of frames and video quality to have a fair comparison. Held *et al.* used the MViT [17, 36] video encoder to extract spatio-temporal features. As SoccerNet-MVFoul contains multiple views for each action, they ag-

Feat. extr.	Pooling	Type of Foul		Offence Severity	
		Acc.	BA.	Acc.	BA.
ResNet [24]	Mean	0.30	0.27	0.34	0.25
ResNet [24]	Max	0.32	0.27	0.32	0.24
R(2+1)D [24]	Mean	0.32	0.34	0.34	0.30
R(2+1)D [24]	Max	0.34	0.33	0.39	0.31
VARS-MViT [24]	Mean	0.44	<b>0.40</b>	0.38	0.31
VARS-MViT [24]	Max	0.45	0.39	0.43	0.34
CLIP-L/14	Single-view	<b>0.51</b>	0.39	0.52	<b>0.35</b>
X-VARS	Single-view	/	/	<b>0.62</b>	<b>0.35</b>

Table 4. **Multi-task classification.** We compare the multi-task classification accuracy of the fine-tuned CLIP-L/14 and X-VARS (fine-tuned CLIP + LLM) to the performance obtained by Held *et al.* [24]. We obtain state-of-the-art performances for three of the four metrics while using a single view instead of multi-views. X-VARS enhances the classification accuracy of offence and severity by 19% compared to the previous SOTA. **Acc.** stands for Accuracy and **BA.** stands for balanced accuracy.

gregate the features of the different views by mean or max pooling. In this work, we fine-tune CLIP on a single view and evaluate it on the same actions, using only a single view instead of the multi-views. Despite fewer views, CLIP outperforms the previous SOTA in three of the four metrics, especially enhancing foul and severity classification, with an increase of 9% in accuracy. To determine the recognition performance of X-VARS, we asked X-VARS for each video clip, if it was a foul or not, and its corresponding severity. We then asked ChatGPT-3.5 to extract the classification predictions from the generated explanations. Finally, comparing these predictions of X-VARS to the ground truth, we observe a significant performance increase to 62% accuracy in determining whether a foul occurred and its severity compared to CLIPs predictions. Hence, X-VARS outperforms the previous state-of-the-art VARS (MViT+Max Pool) system [24] by 19%. However, since most of the explanations of X-VARS do not explicitly indicate the type of foul, it is not possible to accurately extract it from the explanations. For this reason, we were not able to evaluate the accuracy of X-VARS in determining the type of foul.

**Influence of the video tokens.** We investigated if the LLM simply generates its answers based on the multi-task predictions that we give as input to the LLM or if it also considers the video tokens. To test this hypothesis, we use the X-VARS prediction obtained in the previous section, and we compare it to the CLIPs prediction provided as input. Interestingly, X-VARS did not simply replicate the CLIP prediction as it only agreed on 76% of the cases. This result shows that X-VARS, throughout its training, developed the ability to re-evaluate the multi-task predictions and understand that they are not always reliable. Consequently, X-VARS does not only rely on the text predictions for its answers but also incorporates information from the video tokens.

### CLIPs classification predictions *vs.* no text predictions.

To validate our new training paradigm, we compared the quality of our X-VARS trained with classification prediction as additional text input against training it only with video features. To compare the two models, we randomly selected a set of 40 video clips with a uniform distribution of various types of actions and severities. By qualitatively analyzing the results on the selected set, both models generate similar outcomes for most of the video clips. The main difference occurs for less frequent types of actions and severities. For instance, X-VARS without predictions fails to predict a single “No foul” instance and achieves a balanced accuracy of only 29%, while X-VARS obtains a 6% higher balanced accuracy. Figure 4a shows a clip of a defender holding his opponent, an underrepresented action in our dataset. X-VARS without predictions incorrectly predicts: “*No card, as the defender pushed the attacker in the back with low intensity during the fight for the ball, without any risk of injury*”. On the other hand, X-VARS with predictions provides an accurate explanation: “*No card, as the defender is holding onto the attacker’s jersey without it being unsportsmanlike and without any risk of injury*”. Throughout our testing, we encountered several instances where X-VARS with prediction tokens aligned more closely with the ground truth, especially for underrepresented actions. These results show the effectiveness of our new training paradigm in achieving higher accuracy and more accurate explanations.

## 6. Conclusion

In this work, we investigated the potential of using LLMs to enhance transparency and explainability within decision-making processes. We proposed X-VARS, a multi-modal language model, which can perform a multitude of tasks, including video description, question answering, video action recognition, and conducting meaningful conversations based on video content. X-VARS achieves state-of-the-art performance in determining whether a foul and in assessing the corresponding severity. The qualitative results and human study underscore the exceptional capabilities of X-VARS in explaining its decision, indicating its potential to enhance football refereeing by providing accurate decisions and explanations.

**Acknowledgement.** This work was partly supported by the King Abdullah University of Science and Technology (KAUST) Office of Sponsored Research through the Visual Computing Center (VCC) funding and the SDAIA-KAUST Center of Excellence in Data Science and Artificial Intelligence (SDAIA-KAUST AI). J. Held and A. Cioppa are funded by the F.R.S.-FNRS. The present research benefited from computational resources made available on Lucia, the Tier-1 supercomputer of the Walloon Region, infrastructure funded by the Walloon Region under the grant agreement n°1910247.

## References

- [1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Monteiro Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. *arXiv*, abs/2204.14198, 2022. [1](#), [2](#)
- [2] Adrià Arbués Sangüesa, Adrià Martín, Javier Fernández, Coloma Ballester, and Gloria Haro. Using player's body-orientation to model pass feasibility in soccer. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3875–3884, Seattle, WA, USA, Jun. 2020. Inst. Electr. Electron. Eng. (IEEE). [2](#)
- [3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hessei, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Sutskever Ilya, and Dario Amodei. Language models are few-shot learners. *arXiv*, abs/2005.14165, 2020. [1](#), [2](#), [3](#)
- [4] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. *arXiv*, abs/2102.08981, 2021. [3](#)
- [5] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. ShareGPT4V: Improving large multi-modal models with better captions. *arXiv*, abs/2311.12793, 2023. [2](#)
- [6] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, and et al. PaLM: Scaling language modeling with pathways. *arXiv*, abs/2204.02311, 2022. [1](#), [2](#)
- [7] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Scaling up SoccerNet with multi-view spatial localization and re-identification. *Sci. Data*, 9(1):1–9, Jun. 2022. [2](#)
- [8] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Bernard Ghanem, Marc Van Droogenbroeck, Rikke Gade, and Thomas B. Moeslund. A context-aware loss function for action spotting in soccer videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 13123–13133, Seattle, WA, USA, Jun. 2020. Inst. Electr. Electron. Eng. (IEEE). [2](#)
- [9] Anthony Cioppa, Adrien Deliège, Silvio Giancola, Floriane Magera, Olivier Barnich, Bernard Ghanem, and Marc Van Droogenbroeck. Camera calibration and player localization in SoccerNet-v2 and investigation of their representations for action spotting. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, CVsports, pages 4532–4541, Nashville, TN, USA, Jun. 2021. [2](#)
- [10] Anthony Cioppa, Adrien Deliège, Maxime Istasse, Christophe De Vleeschouwer, and Marc Van Droogenbroeck. ARTHuS: Adaptive real-time human segmentation in sports through online distillation. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, CVsports, pages 2505–2514, Long Beach, CA, USA, Jun. 2019. Inst. Electr. Electron. Eng. (IEEE). [2](#)
- [11] Anthony Cioppa, Silvio Giancola, Adrien Deliège, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. SoccerNet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, CVsports, pages 3490–3501, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). [2](#)
- [12] Anthony Cioppa, Silvio Giancola, Vladimir Somers, Floriane Magera, Xin Zhou, Hassan Mkhallati, Adrien Deliège, Jan Held, Carlos Hinojosa, Amir M. Mansourian, Pierre Miralles, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Abdullah Kamal, Adrien Maglo, Albert Clapés, Amr Abdelaziz, Artur Xarles, Astrid Orcesi, Atom Scott, Bin Liu, Byoungkwon Lim, Chen Chen, Fabian Deuser, Feng Yan, Fufu Yu, Gal Shitrit, Guanshuo Wang, Gyusik Choi, Hankyul Kim, Hao Guo, Hasby Fahrudin, Hidenari Koguchi, Håkan Ardö, Ibrahim Salah, Ido Yerushalmey, Iftikar Muhammad, Ikuma Uchida, Ishay Be'ery, Jaonary Rabarisoa, Jeongae Lee, Jiajun Fu, Jianqin Yin, Jinghang Xu, Jongho Nang, Julien Denize, Junjie Li, Junpei Zhang, Juntae Kim, Kamil Synowiec, Kenji Kobayashi, Kexin Zhang, Konrad Habel, Kota Nakajima, Licheng Jiao, Lin Ma, Lizhi Wang, Luping Wang, Menglong Li, Mengying Zhou, Mohamed Nasr, Mohamed Abdelwahed, Mykola Liashuh, Nikolay Falaleev, Norbert Oswald, Qiong Jia, Quoc-Cuong Pham, Ran Song, Romain Héroult, Rui Peng, Ruilong Chen, Ruixuan Liu, Ruslan Baikulov, Ryuto Fukushima, Sergio Escalera, Seungcheon Lee, Shimin Chen, Shouhong Ding, Taiga Someya, Thomas B. Moeslund, Tianjiao Li, Wei Shen, Wei Zhang, Wei Li, Wei Dai, Weixin Luo, Wending Zhao, Wenjie Zhang, Xinquan Yang, Yambiao Ma, Yeeun Joo, Yingsen Zeng, Yiyang Gan, Yongqiang Zhu, Yujie Zhong, Zheng Ruan, Zhiheng Li, Zhijian Huang, and Ziyu Meng. SoccerNet 2023 challenges results. *arXiv*, abs/2309.06006, 2023. [2](#)
- [13] Adrien Deliège, Anthony Cioppa, Silvio Giancola, Meisam J. Seikavandi, Jacob V. Dueholm, Kamal Nasrollahi, Bernard Ghanem, Thomas B. Moeslund, and Marc Van Droogenbroeck. SoccerNet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, CVsports, pages 4508–4519, Nashville, TN, USA, Jun. 2021. [2](#)
- [14] Julien Denize, Mykola Liashuh, Jaonary Rabarisoa, Astrid Orcesi, and Romain Héroult. COMEDIAN: Self-supervised learning and knowledge distillation for action spotting using transformers. *arXiv*, abs/2309.01270, 2023. [2](#)

- [15] Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient finetuning of quantized LLMs. *arXiv*, abs/2305.14314, 2023. 6
- [16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv*, abs/1810.04805, 2018. 1, 2
- [17] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *IEEE Int. Conf. Comput. Vis. (ICCV)*, pages 6804–6815, Montréal, Can., Oct. 2021. Inst. Electr. Electron. Eng. (IEEE). 7
- [18] Jiale Fang, Calvin Yeung, and Keisuke Fujii. Foul prediction with estimated poses from soccer broadcast video. *arXiv*, abs/2402.09650, 2024. 2
- [19] FIFA. Semi automated offside technology. <https://www.fifa.com/technical/football-technology/football-technologies-and-innovations-at-the-fifa-world-cup-2022/semi-automated-offside-technology>, 2023. 2
- [20] Sushant Gautam, Cise Midoglu, Saeed Shafiee Sabet, Dinesh Baniya Kshatri, and Pål Halvorsen. Soccer game summarization using audio commentary, metadata, and captions. *Proceedings of the 1st Workshop on User-centric Narrative Summarization of Long Videos*, Oct. 2022. 2
- [21] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. SoccerNet: A scalable dataset for action spotting in soccer videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 1792–179210, Salt Lake City, UT, USA, Jun. 2018. Inst. Electr. Electron. Eng. (IEEE). 2
- [22] Silvio Giancola, Anthony Cioppa, Adrien Delière, Floriane Magera, Vladimir Somers, Le Kang, Xin Zhou, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Abdulrahman Darwish, Adrien Maglo, Albert Clapés, Andreas Luyts, Andrei Boiarov, Artur Xarles, Astrid Orcesi, Avijit Shah, Baoyu Fan, Bharath Comandur, Chen Chen, Chen Zhang, Chen Zhao, Chengzhi Lin, Cheuk-Yiu Chan, Chun Chuen Hui, Dengjie Li, Fan Yang, Fan Liang, Fang Da, Feng Yan, Fufu Yu, Guanshuo Wang, H. Anthony Chan, He Zhu, Hongwei Kan, Jiaming Chu, Jianming Hu, Jianyang Gu, Jin Chen, João V. B. Soares, Jonas Theiner, Jorge De Corte, José Henrique Brito, Jun Zhang, Junjie Li, Junwei Liang, Leqi Shen, Lin Ma, Lingchi Chen, Miguel Santos Marques, Mike Azatov, Nikita Kasatkina, Ning Wang, Qiong Jia, Quoc Cuong Pham, Ralph Ewerth, Ran Song, Rengang Li, Rikke Gade, Ruben Debien, Runze Zhang, Sangrok Lee, Sergio Escalera, Shan Jiang, Shigeyuki Odashima, Shimin Chen, Shoichi Massui, Shouhong Ding, Sin-wai Chan, Siyu Chen, Tallal El-Shabrawy, Tao He, Thomas B. Moeslund, Wan-Chi Siu, Wei Zhang, Wei Li, Xiangwei Wang, Xiao Tan, Xiaochuan Li, Xiaolin Wei, Xiaoqing Ye, Xing Liu, Xinying Wang, Yandong Guo, Yaqian Zhao, Yi Yu, Yingying Li, Yue He, Yujie Zhong, Zhenhua Guo, and Zhiheng Li. SoccerNet 2022 challenges results. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 75–86, Lisbon, Port., Oct. 2022. ACM. 2
- [23] Silvio Giancola, Anthony Cioppa, Julia Georgieva, Johsan Billingham, Andreas Serner, Kerry Peek, Bernard Ghanem, and Marc Van Droogenbroeck. Towards active learning for action spotting in association football videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 5098–5108, Vancouver, Can., Jun. 2023. Inst. Electr. Electron. Eng. (IEEE). 2
- [24] Jan Held, Anthony Cioppa, Silvio Giancola, Abdullah Hamdi, Bernard Ghanem, and Marc Van Droogenbroeck. VARS: Video assistant referee system for automated soccer decision making from multiple views. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 5086–5097, Vancouver, Can., Jun. 2023. Inst. Electr. Electron. Eng. (IEEE). 2, 6, 7, 8
- [25] Jordan Hoffmann, Sébastien Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karen Simonyan, Erich Elsen, Jack W. Rae, Oriol Vinyals, and Laurent Sifre. Training compute-optimal large language models. *arXiv*, abs/2203.15556, 2022. 1, 2
- [26] Yutaro Honda, Rei Kawakami, Ryota Yoshihashi, Kenta Kato, and Takeshi Naemura. Pass receiver prediction in soccer using video and players’ trajectories. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3502–3511, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2
- [27] James Hong, Haotian Zhang, Michaël Gharbi, Matthew Fisher, and Kayvon Fatahalian. Spotting temporally precise, fine-grained events in video. In *Eur. Conf. Comput. Vis. (ECCV)*, volume 13695 of *Lect. Notes Comput. Sci.*, pages 33–51, Tel Aviv, Israël, 2022. Springer Nat. Switz. 2
- [28] 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. *arXiv*, abs/2106.09685, 2021. 6
- [29] Shaohan Huang, Li Dong, Wenhui Wang, Yaru Hao, Saksham Singhal, Shuming Ma, Tengchao Lv, Lei Cui, Owais Khan Mohammed, Barun Patra, Qiang Liui, Kriti Aggarwal, Zewen Chi, Johan Björck, Vishrav Chaudhary, Subhajit Som, Xia Song, and Furu Wei. Language is not all you need: Aligning perception with language models. *arXiv*, abs/2302.14045, 2023. 2
- [30] IFAB. Laws of the game. Technical report, The International Football Association Board, Zurich, Switzerland, 2022. 3, 6
- [31] Maxime Istasse, Vladimir Somers, Pratheeban Elancheliyan, Jaydeep De, and Davide Zambrano. DeepSportradar-v2: A multi-sport computer vision dataset for sport understandings. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 23–29, Ottawa, Ontario, Can., Oct. 2023. ACM. 2
- [32] Yudong Jiang, Kaixu Cui, Leilei Chen, Canjin Wang, and Changliang Xu. SoccerDB: A large-scale database for comprehensive video understanding. In *Int. ACM Work. Multi-*

- media Content Anal. Sports (MMSports)*, page 1–8, Seattle, WA, USA, Oct. 2020. ACM. [2](#)
- [33] Himabindu Lakkaraju, Dylan Slack, Yuxin Chen, Chenhao Tan, and Sameer Singh. Rethinking explainability as a dialogue: A practitioner’s perspective. *arXiv*, abs/2202.01875, 2022. [2](#)
- [34] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L. Berg. TVQA: Localized, compositional video question answering. *arXiv*, abs/1809.01696, 2018. [3](#)
- [35] Haopeng Li, Andong Deng, Qihong Ke, Jun Liu, Hossein Rahmani, Yulan Guo, Bernt Schiele, and Chen Chen. Sports-QA: A large-scale video question answering benchmark for complex and professional sports. *arXiv*, abs/2401.01505, 2024. [3](#)
- [36] Yanghao Li, Chao-Yuan Wu, Haoqi Fan, Karttikeya Mangalam, Bo Xiong, Jitendra Malik, and Christoph Feichtenhofer. MViTv2: Improved multiscale vision transformers for classification and detection. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, pages 4794–4804, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). [7](#)
- [37] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *arXiv*, abs/2304.08485, 2023. [2](#), [3](#), [6](#)
- [38] Katja Ludwig, Julian Lorenz, Robin Schön, and Rainer Lienhart. All keypoints you need: Detecting arbitrary keypoints on the body of triple, high, and long jump athletes. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 5179–5187, Vancouver, Can., Jun. 2023. Inst. Electr. Electron. Eng. (IEEE). [2](#)
- [39] Scott Lundberg and Su-In Lee. A unified approach to interpreting model predictions. *arXiv*, abs/1705.07874, 2017. [2](#)
- [40] Chenyang Lyu, Minghao Wu, Longyue Wang, Xinting Huang, Bingshuai Liu, Zefeng Du, Shuming Shi, and Zhaopeng Tu. Macaw-LLM: Multi-modal language modeling with image, audio, video, and text integration. *arXiv*, abs/2306.09093, 2023. [6](#)
- [41] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-ChatGPT: Towards detailed video understanding via large vision and language models. *arXiv*, abs/2306.05424, 2023. [1](#), [2](#), [3](#), [4](#), [6](#), [7](#)
- [42] Adrien Maglo, Astrid Orcesi, and Quoc-Cuong Pham. Efficient tracking of team sport players with few game-specific annotations. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3460–3470, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). [2](#)
- [43] Amir M. Mansourian, Vladimir Somers, Christophe De Vleeschouwer, and Shohreh Kasaei. Multi-task learning for joint re-identification, team affiliation, and role classification for sports visual tracking. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, page 103–112, Ottawa, Ontario, Can., Oct. 2023. ACM. [2](#)
- [44] Cise Midoglu, Steven Hicks, Vajira Thambawita, Tomas Kupka, and Pål Halvorsen. MMSys’22 grand challenge on AI-based video production for soccer. In *ACM Multimedia Systems Conference (MMSys)*, pages 1–6, Athlone, Ireland, Jun. 2022. [2](#)
- [45] Cise Midoglu, Saeed Shafiee Sabet, Mehdi Houshangmand Sarkhoosh, Mohammad Majidi, Sushant Gautam, Håkon Maric Solberg, Tomas Kupka, and Pål Halvorsen. Ai-based sports highlight generation for social media. *Proceedings of the 3rd Mile-High Video Conference on zzz*, Feb. 2024. [2](#)
- [46] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. HowTo100M: Learning a text-video embedding by watching hundred million narrated video clips. *arXiv*, abs/1906.03327, 2019. [3](#)
- [47] Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. SoccerNet-caption: Dense video captioning for soccer broadcasts commentaries. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 5074–5085, Vancouver, Can., Jun. 2023. Inst. Electr. Electron. Eng. (IEEE). [2](#), [3](#)
- [48] Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and et al. Training language models to follow instructions with human feedback. *arXiv*, abs/2203.02155, 2022. [2](#)
- [49] Luca Pappalardo, Paolo Cintia, Alessio Rossi, Emanuele Massucco, Paolo Ferragina, Dino Pedreschi, and Fosca Giannotti. A public data set of spatio-temporal match events in soccer competitions. *Sci. Data*, 6(1):1–15, Oct. 2019. [2](#)
- [50] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *Int. Conf. Mach. Learn. (ICML)*, pages 8748–8763, Jul. 2021. [2](#), [4](#)
- [51] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. “why should i trust you??: Explaining the predictions of any classifier. *arXiv*, abs/1602.04938, 2016. [2](#)
- [52] Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, Jonathan Tow, Alexander M. Rush, Stella Biderman, Albert Webson, Pawan Sasanka Ammanamanchi, Thomas Wang, Benoît Sagot, Niklas Muennighoff, Albert Villanova del Moral, Olatunji Ruwase, Rachel Bawden, Stas Bekerman, Angelina McMillan-Major, Iz Beltagy, Huu Nguyen, Lucile Saulnier, Samson Tan, Pedro Ortiz Suarez, Victor Sanh, Hugo Laurençon, Yacine Jernite, Julien Lauvay, Margaret Mitchell, Colin Raffel, Aaron Gokaslan, Adi Simhi, Aitor Soroa, Alham Fikri Aji, Amit Alfassy, Anna Rogers, Ariel Kreisberg Nitzav, Canwen Xu, Chenghao Mou, Chris Emezue, Christopher Klamm, Colin Leong, Daniel van Strien, David Ifeoluwa Adelani, Dragomir Radev, Eduardo González Ponferrada, Efrat Levkovizh, Ethan Kim, Eyal Bar Natan, Francesco De Toni, Gérard Dupont, Germán Kruszewski, Giada Pistilli, Hady Elsahar, Hamza Benyamina, Hieu Tran, Ian Yu, Idris Abdulmumin, Isaac Johnson, Itziar Gonzalez-Dios, Javier de la Rosa, Jenny Chim, Jesse Dodge, Jian Zhu, Jonathan Chang, Jörg Frohberg, Joseph Tobing, Joydeep Bhattacharjee, Khalid Almubarak,

- Kimbo Chen, Kyle Lo, Leandro Von Werra, Leon Weber, Long Phan, Loubna Ben allal, Ludovic Tanguy, Manan Dey, Manuel Romero Muñoz, Maraim Masoud, María Grandury, and et al. BLOOM: A 176B-parameter open-access multi-lingual language model. *arXiv*, abs/2211.05100, 2022. 2
- [53] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, and et al. LAION-5b: An open large-scale dataset for training next generation image-text models. *arXiv*, abs/2210.08402, 2022. 3
- [54] Atom Scott, Ikuma Uchida, Masaki Onishi, Yoshinari Kameda, Kazuhiro Fukui, and Keisuke Fujii. SoccerTrack: A dataset and tracking algorithm for soccer with fish-eye and drone videos. In *IEEE/CVF Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW)*, pages 3568–3578, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2
- [55] Ramprasaath R. Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-CAM: Visual explanations from deep networks via gradient-based localization. *arXiv*, abs/1610.02391, 2016. 2
- [56] Karolina Seweryn, Gabriel Cheć, Szymon Łukasik, and Anna Wróblewska. Improving object detection quality in football through super-resolution techniques. *arXiv*, abs/2402.00163, 2024. 2
- [57] Karolina Seweryn, Anna Wróblewska, and Szymon Łukasik. Survey of action recognition, spotting and spatio-temporal localization in soccer – current trends and research perspectives. *arXiv*, abs/2309.12067, 2023. 2
- [58] João V. B. Soares and Avijit Shah. Action spotting using dense detection anchors revisited: Submission to the SoccerNet challenge 2022. *arXiv*, abs/2206.07846, 2022. 2
- [59] João V. B. Soares, Avijit Shah, and Topojoy Biswas. Temporally precise action spotting in soccer videos using dense detection anchors. In *IEEE Int. Conf. Image Process. (ICIP)*, pages 2796–2800, Bordeaux, France, Oct. 2022. Inst. Electr. Electron. Eng. (IEEE). 2
- [60] Vladimir Somers, Christophe De Vleeschouwer, and Alexandre Alahi. Body part-based representation learning for occluded person Re-Identification. In *IEEE/CVF Winter Conf. Appl. Comput. Vis. (WACV)*, pages 1613–1623, Waikoloa, HI, USA, Jan. 2023. Inst. Electr. Electron. Eng. (IEEE). 2
- [61] Alessandro Suglia, José Lopes, Emanuele Bastianelli, Andrea Vanzo, Shubham Agarwal, Malvina Nikandrou, Lu Yu, Ioannis Konstas, and Verena Rieser. Going for GOAL: A resource for grounded football commentaries. *arXiv*, abs/2211.04534, 2022. 3
- [62] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. MovieQA: Understanding stories in movies through question-answering. *arXiv*, abs/1512.02902, 2015. 3
- [63] Graham Thomas, Rikke Gade, Thomas B. Moeslund, Peter Carr, and Adrian Hilton. Computer vision for sports: current applications and research topics. *Comput. Vis. Image Underst.*, 159:3–18, Jun. 2017. 2
- [64] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and efficient foundation language models. *arXiv*, abs/2302.13971, 2023. 1, 2
- [65] Gabriel Van Zandycke, Vladimir Somers, Maxime Istasse, Carlo Del Don, and Davide Zambrano. DeepSportradar-v1: Computer vision dataset for sports understanding with high quality annotations. In *Int. ACM Work. Multimedia Content Anal. Sports (MMSports)*, pages 1–8, Lisbon, Port., Oct. 2022. ACM. 2
- [66] Renaud Vandeghen, Anthony Cioppa, and Marc Van Droogenbroeck. Semi-supervised training to improve player and ball detection in soccer. In *IEEE Int. Conf. Comput. Vis. Pattern Recognit. Work. (CVPRW), CVsports*, pages 3480–3489, New Orleans, LA, USA, Jun. 2022. Inst. Electr. Electron. Eng. (IEEE). 2
- [67] Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the GDPR. *arXiv*, abs/1711.00399, 2017. 2
- [68] Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. BridgeTower: Building bridges between encoders in vision-language representation learning. *arXiv*, abs/2206.08657, 2022. 2
- [69] Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingen Zhou. mPLUG-Owl2: Revolutionizing multi-modal large language model with modality collaboration. *arXiv*, abs/2311.04257, 2023. 1, 2
- [70] Junqing Yu, Aiping Lei, Zikai Song, Tingting Wang, Hengyou Cai, and Na Feng. Comprehensive dataset of broadcast soccer videos. In *IEEE Conf. Multimedia Inf. Process. Retr. (MIPR)*, pages 418–423, Miami, FL, USA, Apr. 2018. Inst. Electr. Electron. Eng. (IEEE). 2
- [71] Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. MiniGPT-4: Enhancing vision-language understanding with advanced large language models. *arXiv*, abs/2304.10592, 2023. 1, 2, 6