Towards Universal Soccer Video Understanding

Jiayuan Rao[1,2*], Haoning Wu[1,2*], Hao Jiang[3], Ya Zhang[1], Yanfeng Wang[1†], Weidi Xie[1†]
[1]School of Artificial Intelligence, Shanghai Jiao Tong University, China
[2]CMIC, Shanghai Jiao Tong University, China [3]Alibaba Group, China
https://jyrao.github.io/UniSoccer/

Abstract

*As a globally celebrated sport, soccer has attracted widespread interest from fans all over the world. This paper aims to develop a comprehensive multi-modal framework for soccer video understanding. Specifically, we make the following contributions in this paper: (i) we introduce **SoccerReplay-1988**, the largest multi-modal soccer dataset to date, featuring videos and detailed annotations from 1,988 complete matches, with an automated annotation pipeline; (ii) we present an advanced soccer-specific visual encoder, **MatchVision**, which leverages spatiotemporal information across soccer videos and excels in various downstream tasks; (iii) we conduct extensive experiments and ablation studies on event classification, commentary generation, and multi-view foul recognition. MatchVision demonstrates state-of-the-art performance on all of them, substantially outperforming existing models, which highlights the superiority of our proposed data and model. We believe that this work will offer a standard paradigm for sports understanding research.*

"Football is one of the world's best means of communication. It is impartial, apolitical, and universal."
—— Franz Beckenbauer (1945 - 2024)

1. Introduction

Soccer, celebrated worldwide for its significant commercial value, has recently seen great research interest in integrating artificial intelligence (AI) for soccer video understanding. This is primarily motivated by the sport's complexity and the growing demand for enhanced analytics and improved viewing experiences. AI systems facilitate tactical analysis [49], allowing coaches to devise better strategies by uncovering patterns not apparent to the naked eye. In addition, it also supports automated content generation and enriches fan engagement through interactive and personalized content [34, 36, 40]. These capabilities promote a deeper understanding of soccer, simplify content creation, and foster a more engaging experience for fans and professionals.

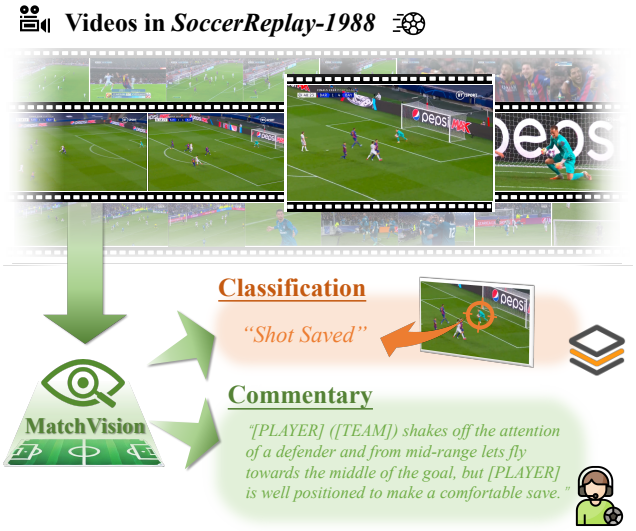


Figure 1. **Overview**. We present **SoccerReplay-1988**, the largest soccer dataset to date, and a powerful soccer-specific visual encoder, **MatchVision**, capable of excelling in various tasks such as event classification and commentary generation.

Existing research in soccer video analysis primarily revolves around the SoccerNet series datasets [7, 10, 15], which comprise 500 full-match videos for benchmarking various tasks, such as action spotting [10, 15] and commentary generation [34, 36, 40]. Despite this extensive coverage, the focus has predominantly been on designing specialized models for task-specific applications, leading to fragmented and incompatible solutions. Such fragmentation underscores the need for a unified framework capable of integrating diverse demands, enabling more holistic and scalable advancements in soccer video understanding.

In this paper, we introduce **SoccerReplay-1988**, the largest and most comprehensive multi-modal soccer video dataset to date, featuring 1,988 complete match videos with

*: These authors contribute equally to this work.
†: Corresponding author.

rich annotations, such as event labels and textual commentaries. This dataset offers a solid foundation for developing advanced soccer understanding models and establishes a challenging new benchmark for the field. Additionally, we have harmonized existing datasets to be compatible with ours, further expanding the available data resources.

Leveraging this dataset, we develop **MatchVision**, an advanced soccer-specific visual encoder tailored for diverse soccer understanding tasks. It employs the cutting-edge visual-language foundation model as the backbone, *e.g.*, SigLIP [58]. We further extend framewise visual features into spatiotemporal representations with temporal attentions [3], by training on diverse visual-language tasks on SoccerReplay-1988, as depicted in Figure 1. As a result, MatchVision exhibits strong adaptability across various tasks, such as event classification and commentary generation, serving as a universal and unified framework for comprehensive soccer video understanding.

To summarize, we make the following contributions in this paper: (i) we construct **SoccerReplay-1988**, the largest and most diverse soccer video dataset to date, featuring videos of 1,988 soccer matches with rich annotations, supported by an automated curation pipeline. This provides a solid foundation for developing robust and comprehensive soccer understanding models; (ii) we present a powerful soccer-specific visual encoder, termed **MatchVision**, which effectively leverages spatiotemporal information in soccer videos, and can adapt to various tasks such as event classification and commentary generation, serving as a unified framework for soccer understanding; (iii) we establish more comprehensive and challenging benchmarks based on our dataset, enabling more professional evaluation of soccer understanding models; (iv) extensive experiments and ablation studies demonstrate the superiority of our data and model across various downstream tasks, achieving state-of-the-art performance on both existing benchmarks and our newly established ones. We believe this work offers a viable paradigm for future sports video understanding.

2. Related Works

Sports Understanding [44] is an evolving field that encompasses multiple research topics and integrates diverse data modalities, covering various tasks such as action spotting [10, 15, 16], commentary generation [34, 37, 40, 51, 52, 57], athlete analysis [41, 55], tactical planning [49], sports health [39], and intelligent refereeing [22, 23]. Furthermore, with the rapid development of multimodal large language models (MLLMs), recent efforts [26, 50, 53, 54] have attempted to build more generalized frameworks to uniformly handle a variety of sports understanding tasks.

Visual-Language Models [1, 27, 28, 38, 58] have exhibited remarkable performance across extensive applications like classification, segmentation, image-text retrieval, and image captioning. Recent efforts have ventured into more challenging video understanding [29, 30, 42, 59, 60] tasks, such as temporal alignment [17, 31], dense captioning [5, 56, 62], and audio description [18–20]. However, these efforts typically focus on general scenarios, limiting their adaptability to specific professional fields. Thus, this paper aims to bridge this gap by advancing visual-language models tailored for comprehensive soccer understanding.

Soccer Game Analysis [9] has primarily focused on tasks such as action spotting [15], replay grounding [10, 61], commentary generation [34, 36, 40], player tracking [8], state reconstruction [43], camera calibration [6, 8, 10] and foul recognition [22, 23], as facilitated by the SoccerNet [7, 10, 14, 15] series datasets, with 500 full-match videos from 2015 to 2017. Unlike existing methods that target designing specific models for distinct tasks, this paper aims to design a unified multi-modal framework that leverages spatiotemporal information within videos, serving as a specialized visual encoder for soccer video understanding.

3. SoccerReplay-1988 Dataset

To establish a solid foundation for soccer understanding, we construct **SoccerReplay-1988**, the largest soccer dataset to date. Here, we first outline our data collection details and an overview of the dataset in Sec. 3.1; followed by elaborating on our automated data curation pipeline in Sec. 3.2; lastly, in Sec. 3.3, we present the data statistics and discussion.

3.1. Dataset Collection

To construct the **SoccerReplay-1988** dataset, we have collected untrimmed, full-match videos from the Internet, encompassing a total of 1,988 matches from six European major soccer leagues and championships[1], spanning the *2014-15* to *2023-24* seasons. For each match, we acquire textual commentaries with second-level timestamps from a sports text live website[2], with part of them annotated with specific event types such as *corner* and *goal*. Additionally, we also incorporate extensive metadata, including detailed background information about the games, players, coaches, referees, and teams, providing a solid foundation for future soccer understanding research.

We partition the SoccerReplay-1988 dataset into train, validation, and test sets, containing 1,488, 250, and 250 full-match videos with diverse and comprehensive annotations, respectively. These sets provide rich training data for downstream tasks, such as event classification and commentary generation, while establishing comprehensive and challenging benchmarks for soccer understanding, as further discussed in subsequent sections.

[1]Premier (England), Laliga (Spain), Bundesliga (Germany), Serie-a (Italy), League-1 (France) and UEFA Champions League.
[2]flashscore.com

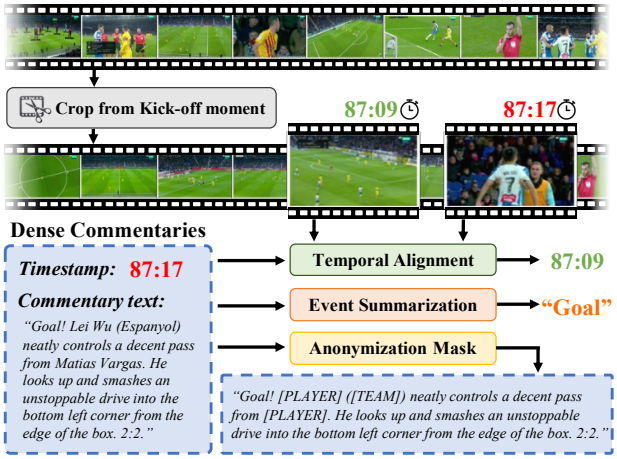


Figure 2. **Automated Data Curation Pipeline.** The collected soccer video data are automatically processed for temporal alignment, event summarization, and anonymization by our curation pipeline.

3.2. Automated Data Curation

Given the potential noise in raw data, such as irrelevant video content, inaccurate timestamps, and incomplete event annotations, we design an automated data curation pipeline, comprising (i) temporal alignment, (ii) event summarization, and (iii) anonymization, as illustrated in Figure 2.

Temporal Alignment. Here, we divide match videos into two halves, each starting at kick-off, and adopt the temporal alignment model from MatchTime [40], to synchronize textual commentary timestamps with those of video frames.

Event Summarization. For samples without event annotations, we leverage LLaMA-3-70B [12] to summarize the events based on textual commentaries. Concretely, we have expanded the event categories from 17 in SoccerNet [10] to 24 types, for finer-grained soccer understanding, for example, categorizing penalties into *scored* and *missed*, and integrating modern soccer regulations like VAR. The resulting 24 event labels include: *'corner'*, *'goal'*, *'injury'*, *'own goal'*, *'penalty'*, *'penalty missed'*, *'red card'*, *'second yellow card'*, *'substitution'*, *'start of game (half)'*, *'end of game (half)'*, *'yellow card'*, *'throw in'*, *'free kick'*, *'saved by goal-keeper'*, *'shot off target'*, *'clearance'*, *'lead to corner'*, *'off-side'*, *'var'*, *'foul (no card)'*, *'statistics and summary'*, *'ball possession'*, and *'ball out of play'*. More details on the used prompts are provided in the **Appendix**.

Anonymization. Similar to [34], we extract all person and team entity names from the metadata of **SoccerReplay-1988**, and replace them in textual commentaries with placeholders, such as "[PLAYER]", "[TEAM]", "[COACH]", and "[REFEREE]", ensuring consistency across tasks.

Moreover, our data curation pipeline can seamlessly extend to existing datasets, converting the SoccerNet series [10, 34] into our unified data format, termed *SoccerNet-pro*. This expansion further enlarges the standardized datasets available for soccer understanding tasks.

	Existing Datasets				
	# Game	Duration(h)	# Event	# Anno.	# Com.
SoccerNet-v1 [15]	500	764	7	6.7k	-
SoccerNet-v2 [10]	500	764	17	110k	-
MatchTime [40]	471	716	14	14k	37k
GOAL [36]	20	25.5	-	-	8.9k
	Our Curated Datasets				
	# Game	Duration(h)	# Event	# Anno.	# Com.
SoccerNet-pro	500	764	24	102k	37k
SoccerReplay-1988	**1,988**	**3,323**	**24**	**150k**	**150k**
Integrated	**2,488**	**4,087**	**24**	**252k**	**187k**

Table 1. **Statistics of Soccer Datasets.** Our SoccerReplay-1988 significantly surpasses existing datasets in both scale and diversity. Here, # Anno. and # Com. refer to the number of event annotations and textual commentaries, respectively.

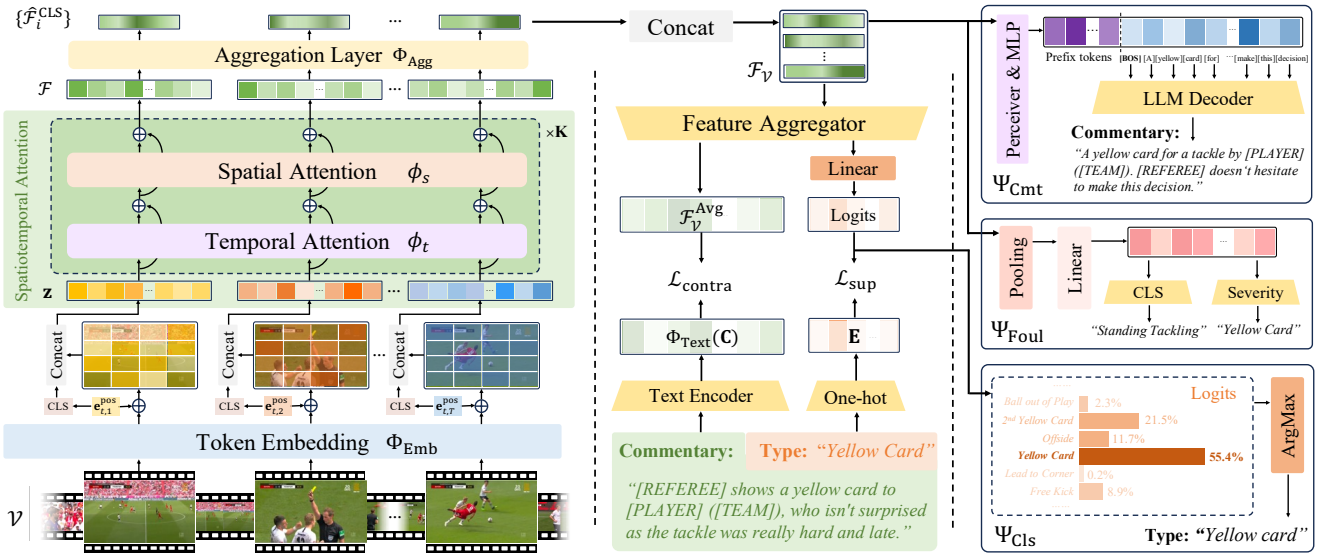
3.3. Statistics & Discussion

Dataset Statistics. As shown in Table 1, our dataset encompasses 3,323 hours of footage from 1,988 soccer matches, with an average duration of 100.3 minutes per match. The videos range in resolution from 360p to 720p and frame rates between 25 and 30 FPS.

For textual annotations, this dataset features approximately 150K commentaries, averaging 76 per match, precisely temporal-aligned by the robust alignment model from MatchTime [40]. These commentaries cover 4,467 unique words, significantly surpassing the 2,873 words in existing datasets [34, 40], greatly enriching textual diversity. Automated event summarization based on these commentaries has yielded about 150K event annotations. Notably, a random sampling of 2% of the data yields 98% manual verification accuracy, ensuring high-quality automated labeling.

SoccerReplay-test Benchmark. To facilitate a more comprehensive evaluation of soccer understanding models, we integrate 250 matches from SoccerReplay-1988 with 50 matches from the curated SoccerNet-pro, establishing **SoccerReplay-test**, a more challenging benchmark for event classification and commentary generation. This benchmark features nearly four times larger than existing datasets and comprises finer-grained event labels, richer textual commentaries, and up-to-date soccer regulations.

Discussion. To summarize, **SoccerReplay-1988** exhibits advancements in three aspects: (i) it is the largest soccer video dataset to date, with nearly four times more videos than existing datasets; (ii) it features more professional and diverse annotations, more suitable for fine-grained and comprehensive soccer understanding tasks; (iii) It employs an automated curation pipeline for annotations and is thus scalable to provide a solid data foundation for future research. All data from SoccerReplay-1988, including videos



(a) Structure of MatchVision (b) MatchVision Pretraining (c) Downstream Tasks

Figure 3. **Overview of MatchVision.** (a) The model architecture and its spatiotemporal feature extraction process; (b) Details of visual encoder pretraining, including supervised classification and video-language contrastive learning; (c) Implementation details of specific heads for various downstream tasks, including commentary generation, foul recognition, and event classification.

and annotations, are open-source for **non-commercial use**.

4. Method

In this paper, we aim to develop a soccer-specific visual encoder, **MatchVision**, tailored for diverse soccer video analysis tasks. We start by outlining our problem formulation in Sec. 4.1. Next, in Sec. 4.2, we detail the architecture of MatchVision. The training procedures are thoroughly discussed in Sec. 4.3. Finally, we describe the configurations for downstream tasks in Sec. 4.4, demonstrating the practical applications and effectiveness of our model.

4.1. Problem Formulation

In this work, we tackle the challenge of analyzing soccer video segments, denoted as $\mathcal{V} \in \mathbb{R}^{T\times 3\times H\times W}$. Our goal is to utilize the visual encoder ($\Phi_{\text{MatchVision}}$) to extract spatiotemporal features from these segments, which are then processed by multiple task-specific heads, formulated as:

$$\mathbf{E}, \mathbf{C}, \mathbf{F} = \Psi(\Phi_{\text{MatchVision}}(\mathcal{V}))$$

Here, $\Psi = \{\Psi_{\text{cls}}, \Psi_{\text{Cmt}}, \Psi_{\text{Foul}}\}$ represents the task-specific heads, with $\mathbf{E}$, $\mathbf{C}$, and $\mathbf{F}$ denoting the output event types, textual commentaries, and foul types, respectively. This unified framework effectively learns relevant spatiotemporal features, and enables seamless integration across various downstream tasks for comprehensive soccer understanding.

4.2. Architecture

MatchVision comprises three key components: (i) Token Embedding, (ii) Spatiotemporal Attention Block, and (iii) Aggregation Layer, as depicted in Figure 3.

Token Embedding. In accordance with the convention in Vision Transformer [11], each frame ($\mathcal{I}_i$) from the video segment ($\mathcal{V} = \{\mathcal{I}_1, \mathcal{I}_2, \cdots, \mathcal{I}_T\}$) is divided into M non-overlapping patches of size $P\times P$ that span the entire frame.

These patches are flattened into vectors ($\mathbf{x}_i^p$), where p and i denote the spatial and temporal positions, respectively. Each vector is transformed via an embedding layer (Φ_{Emb}) into a token vector of size $\mathbb{R}^{1\times D}$, and then added with a spatial position embedding ($\mathbf{e}_{\text{s}}^{\text{pos}} \in \mathbb{R}^{M\times D}$). Subsequently, we concatenate a `[cls]` token along with each frame. Finally, a temporal positional embedding ($\mathbf{e}_{\text{t}}^{\text{pos}} \in \mathbb{R}^{T\times D}$) is added across features of all frames, as formulated below:

$$\mathbf{y}_i = [\mathbf{x}_i^{\text{cls}},\ \Phi_{\text{Emb}}([\mathbf{x}_i^1, \cdots, \mathbf{x}_i^M]) + \mathbf{e}_{\text{s}}^{\text{pos}}]$$
$$\mathbf{z} = [\mathbf{y}_1, \cdots, \mathbf{y}_T] + \mathbf{e}_{\text{t}}^{\text{pos}}$$

Here, $[\cdot,\cdot]$ denotes concatenation, and $\mathbf{y}_i \in \mathbb{R}^{(M+1)\times D}$ represents the frame-wise features. The embedded features ($\mathbf{z}$) will then serve as input for spatiotemporal attention blocks.

Spatiotemporal Attention Block. Similar to TimeSformer [3], we utilize interleaved temporal and spatial attention to integrate spatiotemporal information in soccer videos. Concretely, each spatiotemporal attention block comprises a temporal self-attention layer and a spatial self-attention layer, *i.e.*, $\phi_t(\cdot)$ and $\phi_s(\cdot)$, respectively.

Given a video feature ($\mathbf{z} \in \mathbb{R}^{T\times(M+1)\times D}$), we alternate temporal and spatial attention: temporal attention facilitates interactions among tokens at the same spatial positions across distinct frames, while spatial attention enables interactions among tokens within the same frame. Residual con-

nections are employed in each layer. After passing through a total of $\mathbf{K}$ spatiotemporal attention blocks, the resulting feature ($\mathcal{F}$) captures both intra-frame and inter-frame relationships, *i.e.*, $\mathcal{F} = [\phi_s(\phi_t(\mathbf{z}))]^{\mathbf{K}} \in \mathbb{R}^{T\times(M+1)\times D}$.

Aggregation Layer. To obtain video-level features, we employ an aggregation layer on the frame-wise spatiotemporal features. Specifically, for the i-th frame, we utilize spatial self-attention to aggregate information into its `[cls]` token, denoted as $\hat{\mathcal{F}}_i^{\text{cls}} = \Phi_{\text{Agg}}(\mathcal{F}_i)$. Concatenating the `[cls]` tokens of all frames yields the final video feature ($\mathcal{F}_\mathcal{V}$), that effectively encapsulates spatiotemporal characteristics of soccer video segments, thus enabling it to be applicable for various downstream soccer understanding tasks. This process can be formulated as:

$$\mathcal{F}_\mathcal{V} = \Phi_{\text{MatchVision}}(\mathcal{V}) = [\hat{\mathcal{F}}_1^{\text{cls}}, \cdots, \hat{\mathcal{F}}_T^{\text{cls}}] \in \mathbb{R}^{T\times D}$$

4.3. MatchVision Pretraining

In this part, we aim to pretrain the visual encoder with triplet samples ($\{\mathcal{V}, \mathbf{E}, \mathbf{C}\}$), comprising videos, event labels, and textual commentaries. Concretely, we investigate two distinct pretraining strategies: supervised classification and video-language contrastive learning.

Supervised Classification. One way to pretrain the visual encoder is supervised learning on event classification. To be specific, the extracted visual features ($\mathcal{F}_\mathcal{V}$) are aggregated by a temporal self-attention layer into a learnable `[cls]` token, denoted as $\mathcal{F}_\mathcal{V}^{\text{cls}}$, similar to the spatial-wise aggregation mentioned above. This token is then fed into a linear classifier, and trained with a cross-entropy loss for event classification. The objective is denoted as $\mathcal{L}_{\text{sup}}$.

Video-Language Contrastive Learning. As an alternative, we can also pretrain our visual encoder with video-text contrastive learning. Specifically, we adopt simple average pooling on the video feature to get the aggregated visual feature ($\mathcal{F}_\mathcal{V}^{\text{Avg}}$), and encode the textual commentary ($\mathbf{C}$) with a text encoder (Φ_{Text}). We train the model with sigmoid loss ($\mathcal{L}_{\text{sigmoid}}$), as used in SigLIP [58]. Note that, some video clips may have highly similar commentaries, for example, *'start of the game'*, we treat the commentaries with high similarity in the same batch as positive samples when calculating loss functions. This can be expressed as follows:

$$\mathcal{L}_{\text{contra}} = \mathcal{L}_{\text{sigmoid}}(\mathcal{F}_\mathcal{V}^{\text{Avg}}, \Phi_{\text{Text}}(\mathbf{C}))$$

4.4. Downstream Tasks

After the pretraining mentioned above, MatchVision can now serve as a versatile visual encoder, to map the soccer video segments into visual features ($\mathcal{F}_\mathcal{V}$), for training task-specific heads $\Psi = \{\Psi_{\text{cls}}, \Psi_{\text{Cmt}}, \Psi_{\text{Foul}}\}$ across different downstream tasks, including: (i) event classification, (ii) commentary generation, and (iii) foul recognition.

Event Classification. Similar to supervised classification above, we concatenate a learnable `[cls]` token to aggregate frame-wise visual features via temporal self-attention. This token is then fed into a linear classifier for event classification. The event classification head (Ψ_{cls}) is trained with a cross-entropy loss while freezing the visual encoder.

Commentary Generation. We follow the paradigm in MatchTime [40] to generate anonymized textual commentary for soccer video clips. Concretely, the commentary generation head (Ψ_{Cmt}) employs a Perceiver [25] aggregator to consolidate visual features, which are then projected by a trainable MLP, serving as prefix embeddings for a large language model (LLM). Subsequently, an off-the-shelf LLM decodes these embeddings into textual commentary. We adopt the negative log-likelihood loss, commonly used for auto-regressive next-token prediction.

Foul Recognition. As outlined in [22], the foul recognition task takes multi-view videos from the same scene as inputs, with each sample annotated with a foul class (8 types) and severity (4 levels). We encode these multi-view videos with MatchVision, and aggregate the extracted features into a single feature vector, via either max or average pooling, following the common practice. Subsequently, the foul recognition head (Ψ_{Foul}) employs a shared MLP and two task-specific linear classifiers, to predict foul type and severity, respectively. Similar to event classification, we use the combination of cross-entropy losses on the foul type and severity classification to jointly train Ψ_{Foul}.

Discussion. Pretraining MatchVision on large-scale soccer data equips it with substantial domain-specific knowledge, enabling it to serve as a universal visual encoder adaptable to various downstream soccer understanding tasks.

5. Experiments

This section begins with implementation details in Sec. 5.1; followed by quantitative evaluations across downstream tasks in Sec. 5.2; then, we conduct ablation studies on our SoccerReplay-test benchmark to analyze the effectiveness of the proposed dataset and model in Sec. 5.3; finally, we provide qualitative results for comparison in Sec. 5.4.

5.1. Implementation Details

In our experiments, video segments are sampled at 1FPS around annotated timestamps, capturing a 30-second window for each sample. Frames are resized to 224×224 pixels as inputs. We initialize the embedding layer, spatial attention layers, aggregation layer, and text encoder of MatchVision with pretrained weights from SigLIP Base-16 [58] and adopt LLaMA-3 (8B) [12] as the off-the-shelf LLM decoder for commentary generation. All experiments are conducted on $4\times$ Nvidia H800 GPUs with the AdamW [33] optimizer. Next, we elaborate on the training and evaluation details

Visual Encoder	Dataset			Classification (%)			Commentary				
	SN	MT	SR	Acc.@1	Acc.@3	Acc.@5	B@1	B@4	M	R-L	C
Off-the-shelf Models											
I3D [4]	✗	✗	✗	45.4	82.5	93.2	26.77	5.57	24.17	23.12	18.73
C3D [45]	✗	✗	✗	47.8	85.1	95.0	**28.13**	6.64	24.52	24.23	27.88
ResNet [21]	✗	✗	✗	47.2	84.6	94.4	27.34	6.57	24.72	24.43	27.29
CLIP [38]	✗	✗	✗	48.5	85.5	95.2	26.25	6.51	24.27	24.75	28.17
InternVideo [48]	✗	✗	✗	49.9	**87.0**	**95.9**	27.12	6.54	25.02	24.82	29.90
SigLIP [58]	✗	✗	✗	**50.2**	86.7	95.6	27.85	**6.98**	**25.16**	**25.03**	**31.38**
Pretrain with Supervised Classification											
Baidu [61]	✓	✗	✗	56.4	91.9	97.3	**31.20**	8.88	26.56	26.61	38.93
SigLIP	✓	✗	✗	55.9	89.6	94.9	28.51	7.39	25.96	25.94	35.71
SigLIP	✓	✓	✓	57.9	91.7	97.5	30.95	8.56	25.79	26.17	38.24
MatchVision	✓	✗	✗	82.5	96.6	98.8	29.45	7.92	26.01	26.21	36.15
MatchVision	✓	✓	✓	**84.0**	**97.3**	**99.2**	31.05	**9.06**	**26.94**	**27.93**	**42.20**
Pretrain with Visual-Language Contrastive Learning											
SigLIP	✗	✓	✗	55.4	88.8	97.0	28.72	7.72	25.91	26.17	32.27
SigLIP	✗	✓	✓	66.8	93.7	98.6	30.35	8.12	26.05	26.38	39.41
MatchVision	✗	✓	✗	58.9	89.0	97.1	30.33	7.97	25.48	26.33	33.87
MatchVision	✗	✓	✓	**67.9**	**93.9**	**98.6**	**31.94**	**9.12**	**26.24**	**27.56**	**40.76**
Pretrain with Hybrid Supervised-Contrastive Training											
SigLIP	✓	✓	✗	71.2	94.5	98.7	28.63	7.82	25.74	25.35	34.09
SigLIP	✓	✓	✓	67.1	93.2	98.1	30.71	8.78	26.26	26.74	41.82
MatchVision	✓	✓	✗	76.4	96.0	99.0	30.65	8.33	25.28	26.31	37.23
MatchVision	✓	✓	✓	**80.1**	**97.1**	**99.1**	**33.58**	**9.14**	**26.82**	**28.21**	**44.18**

Table 2. **Quantitative Results on Event Classification and Commentary Generation.** Here, SN, MT, and SR represent finetuning with curated SoccerNet-v2 [10], MatchTime [40], and SoccerReplay-1988, respectively. B, M, R-L, and C refer to BLEU, METEOR, ROUGE-L, and CIDEr metrics, respectively. Within each unit, we denote the best performance in **RED** and the second-best performance in BLUE.

about visual encoder pretraining and downstream tasks.

Visual Encoder Pretraining. For both pretraining strategies, we use a batch size of 40 for 15 epochs. The learning rate for all randomly initialized modules, including the temporal attention layers, aggregator layer, and linear classifier, is set to 1×10^{-4}. Meanwhile, the learning rate for modules initialized with pretrained parameters (including the text encoder) is set to 5×10^{-5}. In contrastive training, we adopt a multi-positive strategy where each textual commentary, based on its event label, considers closely related categories (*e.g. "start of game"* and *"offside"*) as positive samples.

Downstream Tasks. In all downstream tasks, unless otherwise specified, we use the frozen visual encoder for feature extraction and only train the task-specific heads with a learning rate of 1×10^{-4} for 30 epochs. The batch sizes for event classification, commentary generation, and foul recognition are set to 40, 32, and 8, respectively. We adopt specific evaluation metrics for these three tasks: (i) For event classification, we use the top-1/3/5 classification accuracy; (ii) For commentary generation, we employ several commonly-used language evaluation metrics, including BLEU [35], METEOR [2], ROUGE-L [32], and CIDEr [47]; (iii) For foul recognition, we follow the common practice, and report top-1/2 and top-1 accuracy for the foul type and severity classification, respectively.

Benchmarks & Baselines. To ensure fair and reliable comparisons with existing work, we evaluate event classification (24 types) on 100 matches from curated SoccerNet-v2 [10] test set; commentary generation on 49 matches from SN-Caption-test-align benchmark manually aligned in [40]; and foul recognition on MVFoul [22]. We consider various baselines: for the first two tasks, this includes off-the-shelf general visual encoders such as ResNet [21], C3D [45], I3D [4], CLIP [38], SigLIP [58], and InternVideo [48], along with Baidu [61] and SigLIP finetuned with soccer-specific data. For foul recognition, we follow previous work [22, 23] and adopt ResNet [21], R(2+1)D [46], and MViT [13] jointly finetuned with classifiers, as baselines.

5.2. Quantitative Evaluation

As depicted in Table 2, we draw two observations on event classification and commentary generation: (i) visual encoders trained on soccer data substantially outperform off-the-shelf general encoders (ResNet, C3D, I3D, CLIP, and

Visual Encoder			Foul Class		Severity
Backbone	**Train**	**Agg.**	**Acc.@1**	**Acc.@2**	**Acc.@1**
ResNet [21]	✓	Mean	0.31	0.56	0.34
		Max	0.32	0.60	0.32
R(2+1)D [46]	✓	Mean	0.31	0.55	0.34
		Max	0.32	0.56	0.39
MViT [13]	✓	Mean	0.40	0.65	0.38
		Max	**0.47**	0.69	0.43
MatchVision	✗	Mean	0.44	0.53	**0.58**
		Max	0.35	**0.70**	0.46

Table 3. **Quantitative Results on Multi-view Foul Recognition.** Our frozen MatchVision encoder can achieve comparable performance with other jointly finetuned visual encoders.

Pretrain			Classification(%)		
Sup.	**Contra.**	**SR**	**Acc.@1**	**Acc.@3**	**Acc.@5**
✓	✗	✗	62.67	83.00	89.81
✓	✗	✓	**68.03**	**86.90**	**92.38**
✗	✓	✗	46.97	75.53	85.85
✗	✓	✓	57.41	83.13	91.00
✓	✓	✗	56.86	80.30	88.09
✓	✓	✓	63.59	85.21	91.63

Table 4. **Ablations on Event Classification.** We explore the impact of various training settings of our MatchVision encoder on the SoccerReplay-test benchmark. Here, Sup., Contra., and SR refer to supervised classification, visual-language contrastive learning, and the SoccerReplay-1988 dataset, respectively.

InternVideo), underscoring the necessity of building specialized models for soccer understanding; (ii) almost all visual encoders, across all training settings, benefit from **SoccerReplay-1988**, emphasizing the value of constructing large-scale, high-quality data for soccer understanding. Next, we will delve into each task to discuss the results.

Event Classification. With identical training strategies and data, MatchVision considerably outperforms other methods in classification accuracy, demonstrating the superiority of its architecture, which effectively leverages spatiotemporal features within soccer videos. For example, MatchVision achieves a Top-1 accuracy of 82.5%, significantly surpassing SigLIP's 55.9% under the same training conditions. Moreover, models trained via supervised classification excel others, primarily because the pre-training task shares the same objectives as the downstream event classification task.

Commentary Generation. Visual encoders trained with visual-language contrastive learning exhibit better commentary generation performance than those trained with supervised classification, as this strategy better captures correlations between visual and textual features. Additionally, while MatchVision trained solely on SoccerNet slightly underperforms Baidu [61], incorporating SoccerReplay-1988 enables it to outperform on most metrics. This demonstrates that MatchVision can take advantage of large-scale datasets. Finally, a hybrid training approach, starting with supervised classification followed by visual-language contrastive learning, enables MatchVision to achieve optimal performance. This indicates that learning coarse-grained tasks such as classification provides a foundation for fine-grained tasks like commentary generation, and fully leveraging data unlocks the potential of soccer understanding.

Foul Recognition. As demonstrated in Table 3, MatchVision achieves performance comparable to jointly finetuned state-of-the-art methods in foul recognition, even with a frozen visual encoder. This highlights that MatchVision effectively learns substantial knowledge from large-scale soccer data and adapts seamlessly to downstream tasks. Comparisons with additional baselines from the SoccerNet foul recognition challenges [9] are provided in the **Appendix**.

5.3. Ablation Studies

We conduct ablation experiments on event classification and commentary generation using our **SoccerReplay-test** benchmark. These experiments validate the effectiveness of our proposed dataset and model, while establishing a baseline for future evaluations on this benchmark.

Event Classification. We evaluate event classification on 300 matches from our **SoccerReplay-test** benchmark using the MatchVision visual encoder pretrained with various strategies. Features are extracted by MatchVision and processed with a learnable aggregation layer and a linear classifier. The default training set is our curated SoccerNet-pro. As shown in Table 4, integrating SoccerReplay-1988 for training results in significant performance improvements across all pretraining strategies, yielding the significance of our dataset. Additionally, supervised classification outperforms visual-language contrastive learning and hybrid pretraining. This is due to its closer alignment with downstream event classification task, and the scale of event annotations is far larger than that of textual commentaries, further confirming the substantial benefits of data scaling for boosting soccer understanding.

Commentary Generation. With the pretrained MatchVision encoder, we train the commentary generation head on the MatchTime [40] and SoccerReplay-1988 datasets using various training strategies. By default, only the Perceiver [25] aggregation layer and projection layer within the head are trained. For joint training with the LLM decoder, considering computational costs, we incorporate LoRA [24] layers while freezing the original LLM layers. As shown in Table 5, incorporating SoccerReplay-1988 significantly improves performance on all metrics, confirming substantial advantages of our proposed dataset. This performance

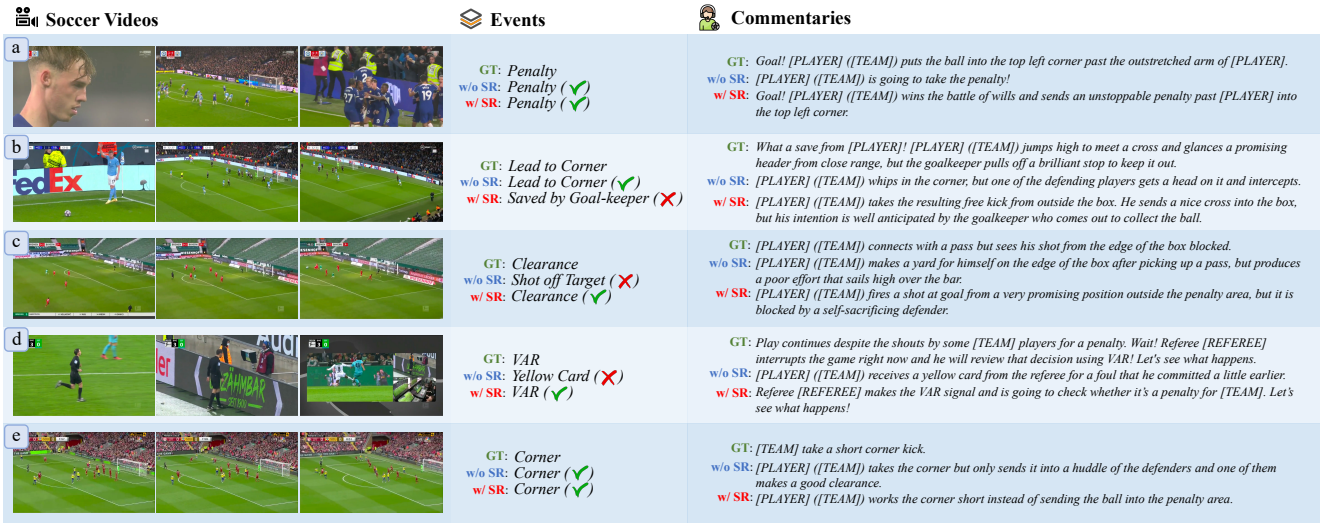


Figure 4. **Qualitative Results for Event Classification and Commentary Generation.** Here, "w/o SR" and "w/ SR" indicate models trained without and with the SoccerReplay-1988 dataset, respectively. Incorporating SoccerReplay-1988 improves event classification accuracy. Moreover, this enriched training data enables the model to demonstrate several advantages in commentary generation: (a) more detailed descriptions, (b) greater linguistic variety, (c) higher event depiction accuracy, (d) better adherence to updated rules, and (e) improved specificity in scenario response.

Trainable		Commentary Metrics				
V	L	B@1	B@4	M	R-L	C
Trained on MatchTime						
✗	✗	21.65	3.27	21.02	17.79	12.90
✓	✗	**27.62**	**7.02**	24.03	**23.51**	30.77
✗	✓	27.04	6.41	24.15	23.88	**31.91**
✓	✓	27.49	6.96	**24.50**	23.33	30.81
Trained on MatchTime & SoccerReplay-1988						
✗	✗	24.17	4.09	20.51	20.70	15.70
✓	✗	28.98	**8.39**	24.45	25.35	**45.85**
✗	✓	27.54	7.76	24.50	24.70	42.79
✓	✓	**29.21**	8.22	**25.25**	**25.54**	43.18

Table 5. **Ablations on Commentary Generation.** We investigate the impact of different training strategies and datasets on MatchVision using the SoccerReplay-test benchmark. 'V' and 'L' denote the visual encoder and the LLM decoder, respectively.

gap also reflects the challenges of our established benchmark, which features diverse vocabulary, richer semantics, and updated soccer rules. Additionally, jointly finetuning the visual encoder and the LLM decoder provides a feasible approach for further improvements.

5.4. Qualitative Comparisons

As depicted in Figure 4, we present qualitative results of MatchVision on the SoccerReplay-test benchmark, comparing models pretrained with and without SoccerReplay-1988. For event classification, incorporating our data improves accuracy, and even in misclassified cases, the results remain contextually relevant. For commentary generation, hybrid training on SoccerReplay-1988 enables MatchVision to produce richer, more detailed textual commentary, reflecting a deeper understanding of soccer dynamics. More qualitative results are available in the **Appendix**.

6. Conclusion

In this paper, we establish a unified, scalable multi-modal framework for soccer understanding. Specifically, we introduce **SoccerReplay-1988**, the largest and most comprehensive soccer video dataset to date, annotated by an automated curation pipeline. This provides a solid foundation for developing soccer understanding models and serves as a more challenging benchmark. Built upon this, we develop **MatchVision**, an advanced soccer-specific visual encoder, which effectively leverages spatiotemporal information within soccer videos and can be applied to various tasks such as event classification and commentary generation. Extensive experiments demonstrate the superiority of our model, with MatchVision achieving state-of-the-art performance on both existing benchmarks and our newly established one. We believe this work will set a viable, universal paradigm for future research in sports understanding.

Acknowledgments

This work is funded by National Key R&D Program of China (No.2022ZD0161400).

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. In *Advances in Neural Information Processing Systems*, pages 23716–23736, 2022. 2

[2] Satanjeev Banerjee and Alon Lavie. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. In *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, pages 65–72, 2005. 6

[3] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *Proceedings of the International Conference on Machine Learning*, page 4, 2021. 2, 4

[4] Joao Carreira and Andrew Zisserman. Quo vadis, action recognition? a new model and the kinetics dataset. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6299–6308, 2017. 6

[5] Tsai-Shien Chen, Aliaksandr Siarohin, Willi Menapace, Ekaterina Deyneka, Hsiang-wei Chao, Byung Eun Jeon, Yuwei Fang, Hsin-Ying Lee, Jian Ren, Ming-Hsuan Yang, and Sergey Tulyakov. Panda-70m: Captioning 70m videos with multiple cross-modality teachers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2

[6] Anthony Cioppa, Adrien Deliege, Floriane Magera, Silvio Giancola, 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 4537–4546, 2021. 2

[7] Anthony Cioppa, Adrien Deliege, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Scaling up soccernet with multi-view spatial localization and re-identification. *Scientific data*, 9(1):355, 2022. 1, 2

[8] Anthony Cioppa, Silvio Giancola, Adrien Deliege, Le Kang, Xin Zhou, Zhiyu Cheng, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-tracking: Multiple object tracking dataset and benchmark in soccer videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 3491–3502, 2022. 2

[9] Anthony Cioppa, Silvio Giancola, Vladimir Somers, Victor Joos, Floriane Magera, Jan Held, Seyed Abolfazl Ghasemzadeh, Xin Zhou, Karolina Seweryn, Mateusz Kowalczyk, Zuzanna Mróz, Szymon Łukasik, Michał Hałoń, Hassan Mkhallati, Adrien Deliège, Carlos Hinojosa, Karen Sanchez, Amir M. Mansourian, Pierre Miralles, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, Bernard Ghanem, Marc Van Droogenbroeck, Adam Gorski, Albert Clapés, Andrei Boiarov, Anton Afanasiev, Artur Xarles, Atom Scott, ByoungKwon Lim, Calvin Yeung, Cristian Gonzalez, Dominic Rüfenacht, Enzo Pacilio, Fabian Deuser, Faisal Sami Altawijri, Francisco Cachón, HanKyul Kim, Haobo Wang, Hyeonmin Choe, Hyunwoo J Kim, Il-Min Kim, Jae-Mo Kang, Jamshid Tursunboev, Jian Yang, Jihwan Hong, Jimin Lee, Jing Zhang, Junseok Lee, Kexin Zhang, Konrad Habel, Licheng Jiao, Linyi Li, Marc Gutiérrez-Pérez, Marcelo Ortega, Menglong Li, Milosz Lopatto, Nikita Kasatkin, Nikolay Nemtsev, Norbert Oswald, Oleg Udin, Pavel Kononov, Pei Geng, Saad Ghazai Alotaibi, Sehyung Kim, Sergei Ulasen, Sergio Escalera, Shanshan Zhang, Shuyuan Yang, Sunghwan Moon, Thomas B. Moeslund, Vasyl Shandyba, Vladimir Golovkin, Wei Dai, WonTaek Chung, Xinyu Liu, Yongqiang Zhu, Youngseo Kim, Yuan Li, Yuting Yang, Yuxuan Xiao, Zehua Cheng, and Zhihao Li. Soccernet 2024 challenges results. *arXiv preprint arXiv:2409.10587*, 2024. 2, 7

[10] Adrien Deliege, 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 4508–4519, 2021. 1, 2, 3, 6

[11] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *Proceedings of the International Conference on Learning Representations*, 2021. 4

[12] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 3, 5

[13] Haoqi Fan, Bo Xiong, Karttikeya Mangalam, Yanghao Li, Zhicheng Yan, Jitendra Malik, and Christoph Feichtenhofer. Multiscale vision transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6824–6835, 2021. 6, 7

[14] Sushant Gautam, Mehdi Houshmand Sarkhoosh, Jan Held, Cise Midoglu, Anthony Cioppa, Silvio Giancola, Vajira Thambawita, Michael A Riegler, Pål Halvorsen, and Mubarak Shah. Soccernet-echoes: A soccer game audio commentary dataset. *arXiv preprint arXiv:2405.07354*, 2024. 2

[15] Silvio Giancola, Mohieddine Amine, Tarek Dghaily, and Bernard Ghanem. Soccernet: A scalable dataset for action spotting in soccer videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1711–1721, 2018. 1, 2, 3

[16] Xiaofan Gu, Xinwei Xue, and Feng Wang. Fine-grained action recognition on a novel basketball dataset. In *International Conference on Acoustics, Speech, and Signal Processing*, pages 2563–2567, 2020. 2

[17] Tengda Han, Weidi Xie, and Andrew Zisserman. Temporal alignment networks for long-term video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2906–2916, 2022. 2

[18] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad: Movie description in

context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2

[19] Tengda Han, Max Bain, Arsha Nagrani, Gul Varol, Weidi Xie, and Andrew Zisserman. Autoad ii: The sequel-who, when, and what in movie audio description. In *Proceedings of the International Conference on Computer Vision*, 2023.

[20] Tengda Han, Max Bain, Arsha Nagrani, Gül Varol, Weidi Xie, and Andrew Zisserman. Autoad iii: The prequel - back to the pixels. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2

[21] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016. 6, 7

[22] 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 *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 5085–5096, 2023. 2, 5, 6

[23] Jan Held, Hani Itani, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. X-vars: Introducing explainability in football refereeing with multimodal large language models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 3267–3279, 2024. 2, 6

[24] 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. In *Proceedings of the International Conference on Learning Representations*, 2022. 7

[25] Andrew Jaegle, Felix Gimeno, Andy Brock, Oriol Vinyals, Andrew Zisserman, and Joao Carreira. Perceiver: General perception with iterative attention. In *Proceedings of the International Conference on Machine Learning*, pages 4651–4664, 2021. 5, 7

[26] Haopeng Li, Andong Deng, Qiuhong 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 preprint arXiv:2401.01505*, 2024. 2

[27] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the International Conference on Machine Learning*, pages 12888–12900, 2022. 2

[28] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the International Conference on Machine Learning*, pages 19730–19742, 2023. 2

[29] Kunchang Li, Yali Wang, Yinan He, Yizhuo Li, Yi Wang, Yi Liu, Zun Wang, Jilan Xu, Guo Chen, Ping Luo, Limin Wang, and Yu Qiao. Mvbench: A comprehensive multi-modal video understanding benchmark. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 22195–22206, 2024. 2

[30] Yanwei Li, Chengyao Wang, and Jiaya Jia. Llama-vid: An image is worth 2 tokens in large language models. In *Proceedings of the European Conference on Computer Vision*, 2024. 2

[31] Zeqian Li, Qirui Chen, Tengda Han, Ya Zhang, Yanfeng Wang, and Weidi Xie. Multi-sentence grounding for long-term instructional video. In *Proceedings of the European Conference on Computer Vision*, 2024. 2

[32] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, 2004. 6

[33] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*, 2019. 5

[34] Hassan Mkhallati, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. Soccernet-caption: Dense video captioning for soccer broadcasts commentaries. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 5074–5085, 2023. 1, 2, 3

[35] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics*, pages 311–318, 2002. 6

[36] Ji Qi, Jifan Yu, Teng Tu, Kunyu Gao, Yifan Xu, Xinyu Guan, Xiaozhi Wang, Bin Xu, Lei Hou, Juanzi Li, et al. Goal: A challenging knowledge-grounded video captioning benchmark for real-time soccer commentary generation. In *Proceedings of the ACM International Conference on Information and Knowledge Management*, pages 5391–5395, 2023. 1, 2, 3

[37] Mengshi Qi, Yunhong Wang, Annan Li, and Jiebo Luo. Sports video captioning via attentive motion representation and group relationship modeling. *IEEE Transactions on Circuits and Systems for Video Technology*, 2019. 2

[38] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the International Conference on Machine Learning*, 2021. 2, 6

[39] Prem N Ramkumar, Bryan C Luu, Heather S Haeberle, Jaret M Karnuta, Benedict U Nwachukwu, and Riley J Williams. Sports medicine and artificial intelligence: a primer. *The American Journal of Sports Medicine*, 50(4): 1166–1174, 2022. 2

[40] Jiayuan Rao, Haoning Wu, Chang Liu, Yanfeng Wang, and Weidi Xie. Matchtime: Towards automatic soccer game commentary generation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2024. 1, 2, 3, 5, 6, 7

[41] Dian Shao, Yue Zhao, Bo Dai, and Dahua Lin. Finegym: A hierarchical video dataset for fine-grained action understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2616–2625, 2020. 2

[42] Yan Shu, Zheng Liu, Peitian Zhang, Minghao Qin, Junjie Zhou, Zhengyang Liang, Tiejun Huang, and Bo Zhao.

Video-xl: Extra-long vision language model for hour-scale video understanding. *arXiv preprint arXiv:2409.14485*, 2024. 2

[43] Vladimir Somers, Victor Joos, Anthony Cioppa, Silvio Giancola, Seyed Abolfazl Ghasemzadeh, Floriane Magera, Baptiste Standaert, Amir M Mansourian, Xin Zhou, Shohreh Kasaei, et al. Soccernet game state reconstruction: End-to-end athlete tracking and identification on a minimap. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 3293–3305, 2024. 2

[44] Graham Thomas, Rikke Gade, Thomas B Moeslund, Peter Carr, and Adrian Hilton. Computer vision for sports: Current applications and research topics. *Computer Vision and Image Understanding*, 159:3–18, 2017. 2

[45] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the International Conference on Computer Vision*, pages 4489–4497, 2015. 6

[46] Du Tran, Heng Wang, Lorenzo Torresani, Jamie Ray, Yann LeCun, and Manohar Paluri. A closer look at spatiotemporal convolutions for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 6450–6459, 2018. 6, 7

[47] Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4566–4575, 2015. 6

[48] Yi Wang, Kunchang Li, Yizhuo Li, Yinan He, Bingkun Huang, Zhiyu Zhao, Hongjie Zhang, Jilan Xu, Yi Liu, Zun Wang, et al. Internvideo: General video foundation models via generative and discriminative learning. *arXiv preprint arXiv:2212.03191*, 2022. 6

[49] Zhe Wang, Petar Veličković, Daniel Hennes, Nenad Tomašev, Laurel Prince, Michael Kaisers, Yoram Bachrach, Romuald Elie, Li Kevin Wenliang, Federico Piccinini, et al. Tacticai: an ai assistant for football tactics. *Nature Communications*, 15(1):1–13, 2024. 1, 2

[50] Dekun Wu, He Zhao, Xingce Bao, and Richard P Wildes. Sports video analysis on large-scale data. In *Proceedings of the European Conference on Computer Vision*, 2022. 2

[51] Zeyu Xi, Ge Shi, Xuefen Li, Junchi Yan, Zun Li, Lifang Wu, Zilin Liu, and Liang Wang. A simple yet effective knowledge guided method for entity-aware video captioning on a basketball benchmark. *Neurocomputing*, 2025. 2

[52] Zeyu Xi, Ge Shi, Haoying Sun, Bowen Zhang, Shuyi Li, and Lifang Wu. Eika: Explicit & implicit knowledge-augmented network for entity-aware sports video captioning. *Expert Systems with Applications*, 2025. 2

[53] Haotian Xia, Zhengbang Yang, Yuqing Wang, Rhys Tracy, Yun Zhao, Dongdong Huang, Zezhi Chen, Yan Zhu, Yuanfang Wang, and Weining Shen. Sportqa: A benchmark for sports understanding in large language models. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics*, 2024. 2

[54] Haotian Xia, Zhengbang Yang, Junbo Zou, Rhys Tracy, Yuqing Wang, Chi Lu, Christopher Lai, Yanjun He, Xun Shao, Zhuoqing Xie, et al. Sportu: A comprehensive sports understanding benchmark for multimodal large language models. In *Proceedings of the International Conference on Learning Representations*, 2025. 2

[55] Jinglin Xu, Yongming Rao, Xumin Yu, Guangyi Chen, Jie Zhou, and Jiwen Lu. Finediving: A fine-grained dataset for procedure-aware action quality assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2949–2958, 2022. 2

[56] Antoine Yang, Arsha Nagrani, Paul Hongsuck Seo, Antoine Miech, Jordi Pont-Tuset, Ivan Laptev, Josef Sivic, and Cordelia Schmid. Vid2seq: Large-scale pretraining of a visual language model for dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 10714–10726, 2023. 2

[57] Huanyu Yu, Shuo Cheng, Bingbing Ni, Minsi Wang, Jian Zhang, and Xiaokang Yang. Fine-grained video captioning for sports narrative. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2018. 2

[58] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the International Conference on Computer Vision*, pages 11975–11986, 2023. 2, 5, 6

[59] Hang Zhang, Xin Li, and Lidong Bing. Video-llama: An instruction-tuned audio-visual language model for video understanding. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2023. 2

[60] Junjie Zhou, Yan Shu, Bo Zhao, Boya Wu, Shitao Xiao, Xi Yang, Yongping Xiong, Bo Zhang, Tiejun Huang, and Zheng Liu. Mlvu: A comprehensive benchmark for multi-task long video understanding. *arXiv preprint arXiv:2406.04264*, 2024. 2

[61] Xin Zhou, Le Kang, Zhiyu Cheng, Bo He, and Jingyu Xin. Feature combination meets attention: Baidu soccer embeddings and transformer based temporal detection. *arXiv preprint arXiv:2106.14447*, 2021. 2, 6, 7

[62] Xingyi Zhou, Anurag Arnab, Shyamal Buch, Shen Yan, Austin Myers, Xuehan Xiong, Arsha Nagrani, and Cordelia Schmid. Streaming dense video captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2024. 2