

SoccerMaster: A Vision Foundation Model for Soccer Understanding

Haolin Yang^{1,2}, Jiayuan Rao¹, Haoning Wu¹, Weidi Xie¹

¹School of Artificial Intelligence, Shanghai Jiao Tong University

²Shanghai Innovation Institute



Figure 1. **SoccerMaster** is a unified soccer-specific vision foundation model that leverages diverse soccer content, including images and videos, to support a wide range of soccer understanding tasks, such as commentary generation, detection, tracking, classification, *etc.*

Abstract

Soccer understanding has recently garnered growing research interest due to its domain-specific complexity and unique challenges. Unlike prior works that typically rely on isolated, task-specific expert models, this work aims to propose a unified model to handle diverse soccer visual understanding tasks, ranging from fine-grained perception (e.g., athlete detection and identification) to high-level semantic reasoning (e.g., event classification). Specifically, our contributions are threefold: (i) we present **SoccerMaster**, the first soccer-specific vision foundation model that unifies diverse tasks within a single framework via **supervised multi-task pretraining**; (ii) we develop an automated data curation pipeline, **SoccerFactory**, to generate scalable spatial annotations, and integrate multiple existing soccer video datasets as a comprehensive pretraining data resource; and (iii) we conduct extensive evaluations demonstrating that **SoccerMaster** consistently outperforms task-specific expert models across diverse downstream tasks, highlighting its breadth and superiority. The data, code, and model will be publicly available at <https://haolinyang-hlyang.github.io/SoccerMaster>.

1. Introduction

Soccer, as one of the world’s most popular sports, enjoys a vast global audience and generates substantial commercial and cultural impact. Such popularity has driven growing research interest in soccer visual understanding [8, 12, 19, 48, 52, 53]. However, the sport’s inherent dynamism, complex rules, and varying broadcast conditions present unique challenges for computer vision systems.

A central dichotomy in soccer understanding lies between *spatial perception* and *semantic reasoning*. The former requires fine-grained geometric precision for tasks such as athlete detection [59, 62], tracking [9, 40, 41], jersey recognition [2, 28], and camera calibration [7, 39]. The latter demands high-level abstraction for action spotting [19], event classification [12], foul recognition [25, 26], and commentary generation [43, 48, 51]. Current paradigms predominantly rely on fragmented, task-specific expert models [9, 26, 51] to address the challenges in isolation. While recent initiatives [53] have attempted to unify these tasks via vision-language alignment, they often neglect dense spatial objectives during pretraining. This omission results in models that can describe *what* is happening but struggle to pinpoint *where* or *who* is involved, leading to a disconnect between spatial perception and semantic reasoning.

To address the aforementioned challenges, we introduce **SoccerMaster**, the first soccer-specific vision foundation model that unifies fine-grained spatial perception and high-level semantic representations within a single framework. As depicted in Fig. 1, SoccerMaster is designed to learn holistic representations through multi-task learning. By simultaneously optimizing for (i) athlete detection and identification, (ii) pitch registration, (iii) event classification, and (iv) vision–language alignment, we enforce a shared representation that captures both the geometry of the pitch and the semantics of the game.

Large-scale training requires massive data. To address the scarcity of dense spatial labels, we develop **SoccerFactory**, an automated data curation pipeline capable of generating high-quality annotations from broadcast footage at scale. By integrating the data curated by SoccerFactory with existing heterogeneous datasets, we construct a comprehensive pretraining resource. This resource balances *breadth* (coverage of diverse events, leagues, and camera styles) with *depth* (dense spatial labels and temporally precise semantics), driving effective multi-task optimization.

We conduct extensive evaluations across a broad spectrum of tasks, spanning both spatial perception (*e.g.*, athlete detection, pitch registration, camera calibration, multiple object tracking) and semantic reasoning (*e.g.*, vision–language alignment, event classification, commentary generation). Our results demonstrate that SoccerMaster substantially outperforms both general-purpose vision foundation models (*e.g.*, SigLIP 2 [61], DINOv3 [58]) and soccer-specific models (MatchVision [53]) across all pre-training tasks with direct evaluation. Moreover, with only lightweight task-specific fine-tuning, our model establishes state-of-the-art or competitive results on downstream applications such as camera calibration [20, 39], multiple object tracking [9], and commentary generation [51]. These findings validate that SoccerMaster has learned universal, transferable representations that effectively generalize to diverse soccer visual understanding scenarios.

Overall, our contributions can be summarized as follows: (i) we introduce **SoccerMaster**, the first soccer-specific vision foundation model that unifies diverse soccer visual understanding tasks within a single framework, serving as a versatile backbone for comprehensive soccer analysis; (ii) we introduce a supervised multi-task pretraining strategy that jointly optimizes fine-grained spatial perception and high-level semantic reasoning tasks on soccer videos, enabling the model to learn multi-granularity representations that capture both local spatial details and global semantic context; (iii) we develop **SoccerFactory**, an automated data curation pipeline to generate large-scale spatial annotations from broadcast footage, and integrate these with various existing soccer video datasets to construct a comprehensive soccer-specific pretraining resource; and (iv) we con-

duct extensive evaluations demonstrating that SoccerMaster can serve as a foundation model for soccer understanding, which achieves state-of-the-art or competitive performance on all downstream tasks through simple fine-tuning.

2. Related Work

Vision foundation models (VFM) [13, 35, 44, 50, 54, 58, 61, 75, 76, 79, 81] have revolutionized computer vision by learning universal visual representations through large-scale pretraining, demonstrating strong generalization across diverse tasks. Their success has inspired applications in various fields, including image captioning [30, 31], 3D perception [65, 77], medical imaging [66, 82], robotics [16], and autonomous driving [17]. However, their exploration in video domains, especially in highly dynamic sports visual understanding, remains relatively nascent.

Sports analysis [60] has gradually become a rapidly evolving field with significant commercial and research value. While the general sports artificial intelligence [29, 67, 71, 72] remains under exploration, recent research on specific sports understanding tasks has achieved significant progress, including action spotting [32], automated scoring [45, 56, 73, 74], commentary generation [67–70], activity recognition [27, 38, 47], tracking [11, 49], and game state analysis [36, 42]. Among these research efforts, soccer, as a globally popular sub-domain with rich spatiotemporal dynamics, has received widespread attention.

Visual soccer understanding involves diverse tasks [10, 21], while early research predominantly focuses on building specific expert models to individually address particular tasks, such as athlete detection [59, 62], tracking [11, 41], jersey number recognition [2, 59], camera calibration [15, 24], event classification [12, 19], commentary generation [43, 48, 51, 78] and foul recognition [25, 26]. This fragmented paradigm inevitably increases the exploration cost in the soccer understanding domain. Although recent work [52] has attempted to leverage multimodal large language models (MLLMs) to handle multiple tasks simultaneously, they only focus on semantic-level tasks such as question-answering. These limitations inspire us to build SoccerMaster, a vision foundation model for the soccer domain that unifies fine-grained spatial perception and high-level semantic reasoning tasks within a single framework.

3. Task Specifications

To develop and evaluate our vision foundation model, we define a set of tasks over a given video clip, $\mathcal{V} = \{\mathcal{I}_1, \dots, \mathcal{I}_T\}$, where $\mathcal{I}_t \in \mathbb{R}^{H \times W \times 3}$ denotes a RGB frame. These tasks involve both fine-grained *spatial perception* and high-level *semantic reasoning*, and are categorized into two groups: **pretraining tasks** that are used to jointly train **SoccerMaster**, including (i) athlete detection and iden-

tification, (ii) pitch registration, (iii) event classification, (iv) vision-language alignment, and **downstream tasks** that demonstrate the model’s generalization ability, including (v) commentary generation, (vi) camera calibration, (vii) multiple object tracking. Details are presented as follows.

Athlete detection and identification. Accurately localizing and identifying athletes is a core task in soccer analysis. For each frame ($\mathcal{I}_t$), we detect athlete instances ($\mathcal{A}_t = \{\mathbf{a}_i \mid i = 1, \dots, N_a\} \in \mathbb{R}^{N_a \times 6}$), where each instance is characterized by $\mathbf{a}_i = (\mathbf{b}_i; \mathbf{r}_i; \mathbf{n}_i)$. Here, $\mathbf{b}_i = (x_i, y_i, w_i, h_i) \in \mathbb{R}^4$ represents the 2D bounding box; $\mathbf{r}_i \in \{\text{goalkeeper, player, referee}\}$ denotes the athlete’s role, and $\mathbf{n}_i \in \mathbb{N}$ indicates the jersey number. Notably, $\mathbf{n}_i$ is set to `null` for goalkeepers and referees, or when occluded.

Pitch registration. Another key spatial perception task lies in detecting field keypoints and lines, which serves as the cornerstone for camera calibration [15, 24]. The objective is to identify and localize structural elements of the soccer pitch, including keypoints (e.g., intersection between side lines) and lines (e.g., middle line and goal crossbar). For each frame ($\mathcal{I}_t$), this task expects to output two sets:

$$\mathcal{K}_t = \{\mathbf{k}_i = (x_i, y_i; t_i) \mid i = 1, \dots, N_k\} \in \mathbb{R}^{N_k \times 3} \quad (1)$$

$$\mathcal{L}_t = \{\mathbf{l}_j = (\mathbf{p}_j^a, \mathbf{p}_j^b; t_j) \mid j = 1, \dots, N_l\} \in \mathbb{R}^{N_l \times 5} \quad (2)$$

where $\mathbf{k}_i$ represents a keypoint with coordinates (x_i, y_i) and type ($t_i \in \mathcal{T}_{\text{keypoint}}$), while $\mathbf{l}_j$ denotes a line segment determined by two unordered endpoints $\mathbf{p}_j^a = (x_j^a, y_j^a)$ and $\mathbf{p}_j^b = (x_j^b, y_j^b)$ with line type ($t_j \in \mathcal{T}_{\text{line}}$). Here, candidate sets $\mathcal{T}_{\text{keypoint}}$ and $\mathcal{T}_{\text{line}}$ follow the definition in [24].

Event classification. As a core semantic reasoning task in soccer understanding, event classification aims to recognize key actions within video segments. Given a video clip ($\mathcal{V}$), the model predicts an event label, $\mathbf{e} \in \mathcal{E}$, where $\mathcal{E}$ denotes the set of 24 event categories defined in SoccerReplay-1988 dataset [53], including *goal*, *corner*, *yellow card*, etc.

Vision-language alignment. This task aims to learn semantic correspondences between video content and textual commentary. Concretely, the objective is to align the video semantic features with the text features extracted by a pre-trained text encoder (e.g. SigLIP 2 [61]), ensuring high similarity between their embeddings in a shared space.

Commentary generation. Building upon vision-language alignment pretraining, commentary generation aims to generate a textual commentary ($\hat{\mathcal{T}}$) of the events and actions occurring in a soccer video clip ($\mathcal{V}$), following [51, 53].

Camera calibration. This task aims to estimate the geometric relationship between the 2D image plane and the 3D soccer field. Formally, given an input frame ($\mathcal{I} \in \mathbb{R}^{H \times W \times 3}$), the task is to predict the camera intrinsics ($\mathbf{K} \in \mathbb{R}^{3 \times 3}$) and extrinsics ($(\mathbf{R}|\mathbf{t})$), where $\mathbf{R} \in \mathbb{R}^{3 \times 3}$ is the rotation matrix and $\mathbf{t} \in \mathbb{R}^3$ is the translation vector.

Multiple object tracking. This task extends athlete detection to establish temporal identity consistency across video frames. Beyond predicting bounding boxes for each athlete in each frame, it assigns a unique identity ($\text{id}_i \in \mathbb{N}$) to each detection, thereby constructing trajectories that maintain consistent identities throughout the video.

Training such a foundation model for diverse soccer understanding tasks demands data spanning both spatial perception and semantic reasoning. While existing datasets cover semantic tasks [12, 51, 53], fine-grained spatial annotations [59] remain limited and costly to scale up.

4. Pretraining Dataset

To address the scarcity of spatial annotations, we develop **SoccerFactory**, an automated data curation pipeline that extracts labels from broadcast footage. By combining these generated annotations with existing resources, we construct a comprehensive dataset that supports unified pretraining for spatial perception and semantic reasoning tasks. This section introduces the automated annotation pipeline in Sec. 4.1, evaluates its quality in Sec. 4.2, and presents the composition of the integrated dataset in Sec. 4.3.

4.1. The SoccerFactory Pipeline

The **SoccerFactory** pipeline consists of three core stages: field registration, tracking, and identification, as well as post-processing refinement, as illustrated in Fig. 2.

Field registration. Following [59], we establish geometric correspondences between image and pitch coordinates through field keypoint and line detectors. The detected keypoints and lines are then processed using the PnL module [24] to estimate camera parameters, enabling projection between image and standardized pitch coordinates. This further allows refining the detected keypoints and lines as training annotations via projection from the canonical pitch.

Tracking and identification. We employ a fine-tuned YOLOv8 [63] to detect athletes, followed by StrongSORT [14] tracking with ReID embeddings extracted by PRTReID [41]. For each detection, we crop the bounding box region and apply Qwen2.5-VL [1] to recognize roles and jersey numbers, which are further filtered by a legibility classifier [28]. Team affiliation is determined through clustering on tracklet-averaged ReID embeddings and pitch coordinates mapped via the estimated camera parameters.

Post-processing refinement. We enhance annotation quality through SAM2-based [54] segmentation refinement to recover missed detections from the detector and correct tracking identity switches. Additionally, we use majority voting across tracklets for jersey numbers and role classifications to improve temporal consistency. To further consolidate fragmented trajectories, we merge short tracklet fragments into longer tracklets by leveraging both ReID embed-

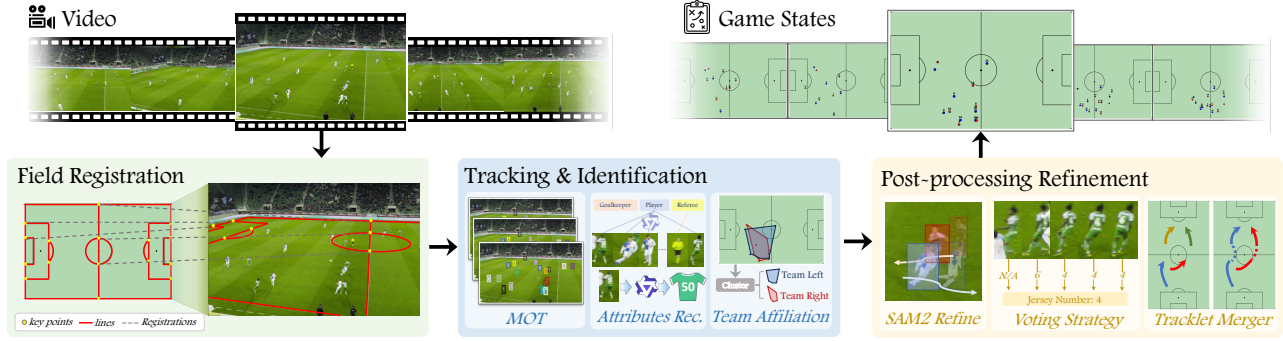


Figure 2. **Automated Data Curation Pipeline.** Our pipeline processes input videos through three stages: (i) field registration establishes geometric correspondences between image and canonical pitch coordinates via keypoint detection; (ii) tracking and identification transforms frames into athlete trajectories through detection, role and team classification, and ReID-based tracking; and (iii) post-processing refinement improves tracking accuracy through SAM2-based segmentation and ensures temporal consistency via majority voting.

Table 1. **Comparison on GSR.** SoccerMaster outperforms the challenge winner KIST-GSR on the SoccerNet-GSR test set. The best and second-best results are **bolded** and underlined.

Method	GS-HOTA	GS-DetA	GS-AssA
Playbox & MIXI	58.1	41.3	81.6
Metricka-Sports	58.2	44.4	76.2
KIST-GSR	<u>61.5</u>	<u>48.5</u>	78.0
Ours	64.1	51.5	<u>79.9</u>

dings and jersey number consistency, similar to [22].

The pipeline generates comprehensive annotations including: athlete bounding boxes with roles, team affiliations, and jersey numbers; soccer field keypoints and lines; camera parameters; and player trajectories in pitch coordinates. These automatically generated annotations enable scalable data production for spatial perception tasks.

4.2. Quality Assessment of SoccerFactory

To assess **SoccerFactory**, we evaluate it on the SoccerNet Game State Reconstruction (GSR) challenge test set [21]. This benchmark covers athlete tracking, role classification, jersey number recognition, camera calibration, and team affiliation. As shown in Tab. 1, our pipeline outperforms state-of-the-art methods with a GS-HOTA score of **64.1**, demonstrating its effectiveness in generating large-scale annotations as a practical alternative to manual labeling.

4.3. Pretraining Dataset Composition

We integrate automatically generated annotations from SoccerFactory with multiple existing soccer datasets to construct a comprehensive, large-scale pretraining dataset, as detailed in Tab. 2. For spatial perception tasks, our pipeline processes main-camera clips from 500 matches in SoccerNet-v2 [12]. To ensure comprehensive field coverage, we exclude broadcast views with incomplete visibility, such as replays, close-ups, and alternative angles. These

Table 2. **Pretraining Dataset Composition.** Our pretraining dataset comprises approximately 7.45M frames across 248.3K video segments, with 2.75M frames for spatial perception tasks and 4.71M frames for semantic reasoning tasks (sampled at 1FPS).

Dataset	# Frames	Pretraining Tasks
<i>Spatial Perception Tasks</i>		
SoccerNet-GSR [59]	42.8K	Athlete Detection and Identification Pitch Registration
SoccerFactory-generated	2.7M	Athlete Detection and Identification Pitch Registration
<i>Semantic Reasoning Tasks</i>		
SoccerNet-v2 [12]	1.6M	Event Classification
MatchTime [51]	7.1K	Event Classification Vision-Language Alignment
SoccerReplay-1988 [53]	3.1M	Event Classification Vision-Language Alignment

generated annotations are combined with manually annotated data from SoccerNet-GSR [59], with tracklet-level jersey numbers converted to frame-level labels via a legibility filter [28] to mitigate occlusion ambiguity. For semantic reasoning tasks, we incorporate large-scale videos (sampled at 1FPS) from SoccerNet-v2 [12], MatchTime [51], and SoccerReplay-1988 [53]. Overall, our pretraining dataset contains 7.45M frames across 248.3K segments (2.75M for spatial perception, 4.71M for semantic reasoning), supporting unified pretraining across diverse soccer tasks. Further details are provided in the **Supplementary Material**.

5. Models

Built on the task definitions in Sec. 3 and the data overview in Sec. 4, this section elaborates on the architecture, training, and evaluation of SoccerMaster, as depicted in Fig. 3. We begin with problem formulations in Sec. 5.1, followed by the model architecture in Sec. 5.2 and multi-task pretraining strategy in Sec. 5.3. Finally, Sec. 5.4 describes how the model is adapted to various downstream tasks.

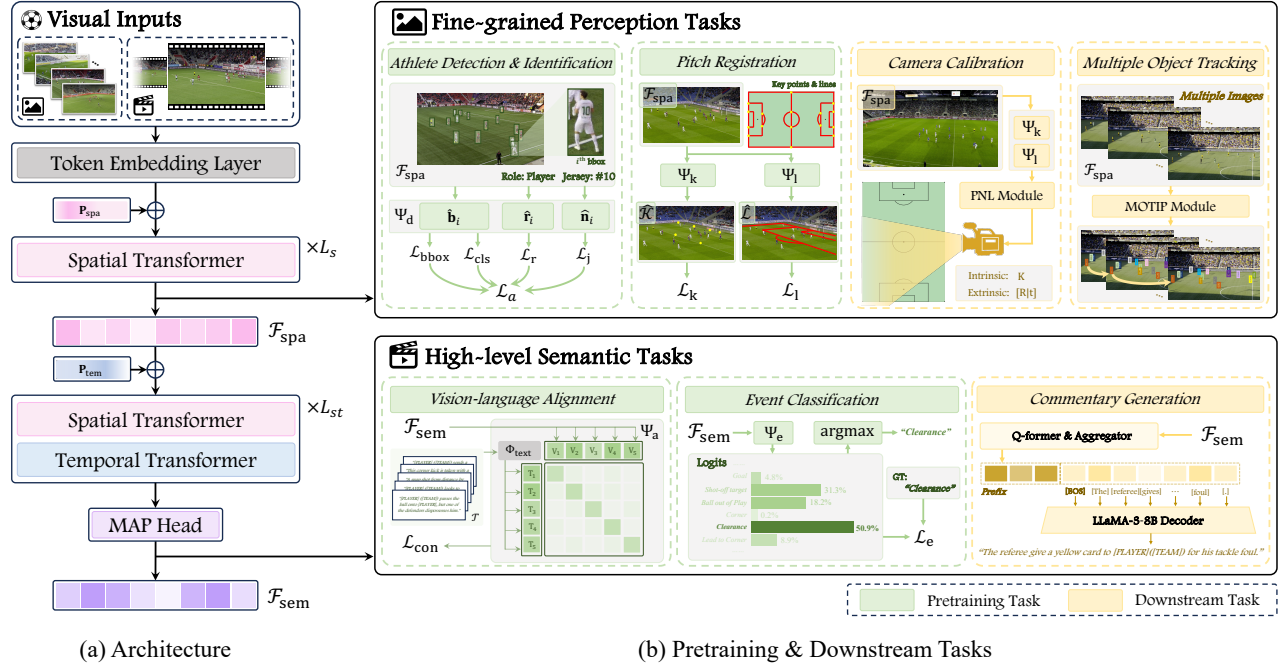


Figure 3. **SoccerMaster Architecture.** (a) The architecture of SoccerMaster, which encodes both soccer videos and images through spatial and temporal attention modules to generate semantically rich representations. (b) The pretraining tasks and downstream adaptations of SoccerMaster across both spatial perception and semantic understanding tasks.

5.1. Problem Formulation

Our soccer vision foundation model, **SoccerMaster**, consists of a visual encoder (Φ_{enc}) and multiple task-specific heads that enable diverse soccer visual understanding capabilities. Given a soccer video segment ($\mathcal{V} = \{\mathcal{I}_1, \dots, \mathcal{I}_T\} \in \mathbb{R}^{T \times 3 \times H \times W}$), our visual encoder (Φ_{enc}) first extracts spatial features ($\mathcal{F}_{\text{spa}}$) and semantic features ($\mathcal{F}_{\text{sem}}$) embedded in the video, which can be formulated as:

$$\mathcal{F} = \{\mathcal{F}_{\text{spa}}, \mathcal{F}_{\text{sem}}\} = \Phi_{\text{enc}}(\mathcal{V}) \quad (3)$$

Leveraging the extracted features ($\mathcal{F}$), our model tackles the aforementioned tasks via corresponding task-specific heads (Ψ_{out}), with the complete set of outputs unified as:

$$\{\mathcal{A}, \{\text{id}_i\}, \mathcal{K}, \mathcal{L}, \mathbf{K}, \mathbf{R}, \mathbf{t}, \mathbf{e}, \hat{\mathcal{T}}\} = \Psi_{\text{out}}(\mathcal{F}) \quad (4)$$

For vision-language alignment, the alignment head (Ψ_a) computes the similarity (s) between video semantic features ($\mathcal{F}_{\text{sem}}$) and text features of textual commentary ($\mathcal{T}$), extracted by a pretrained text encoder (Φ_{text}), denoted as:

$$s = \Psi_a(\mathcal{F}_{\text{sem}}, \Phi_{\text{text}}(\mathcal{T})) \quad (5)$$

5.2. Visual Encoder

Our visual encoder (Φ_{enc}) inherits the ViT [13] structure, consisting of $(L_s + L_{st})$ transformer encoder layers, incorporates a TimeSformer-like [5] design to enable spatiotemporal attention and capture temporal features from videos.

To balance performance and efficiency, we use spatial attention in the first L_s layers and apply spatiotemporal attention only in the final L_{st} layers. The visual feature extraction process comprises three stages: token embedding, spatial encoding, and spatiotemporal encoding, as detailed below.

Token embedding. Given an input video segment ($\mathcal{V} = \{\mathcal{I}_1, \dots, \mathcal{I}_T\} \in \mathbb{R}^{T \times 3 \times H \times W}$), we process each frame following the tokenization procedure of ViT [13]. Specifically, each frame is partitioned into N non-overlapping patches, which are then linearly projected into d -dimensional embeddings and augmented with learnable spatial positional embeddings $\mathbf{P}_{\text{spa}} \in \mathbb{R}^{N \times d}$. This yields the initial token sequence $\mathbf{z}^{(0)} \in \mathbb{R}^{T \times N \times d}$ across all frames, which preserves spatial layout information within each frame.

Spatial encoding. The token sequence ($\mathbf{z}^{(0)}$) first passes through L_s spatial transformer blocks, which process individual frames separately. Each block adopts the standard transformer architecture [13], consisting of multi-head self-attention, LayerNorm, and feedforward network (FFN). For clarity, we focus on describing the attention mechanism, which is the key modification in our design. Given a token ($\mathbf{z}_{t,i}^{(l)}$) at spatial position i and temporal position t in the l -th block, the spatial self-attention can be expressed as:

$$\mathbf{z}_{t,i}^{(l+1)} = \text{SpatialAttn}(\mathbf{z}_{t,i}^{(l)}, \{\mathbf{z}_{t,j}^{(l)}\}_{j=1}^N) \quad (6)$$

Here, each token attends exclusively to all tokens at the same temporal position (t), with no information exchange

across different frames, effectively acting as standard attentions within frames. The output from the L_s -th spatial attention block serves as the extracted **spatial features**, denoted as $\mathcal{F}_{\text{spa}} = \mathbf{z}^{(L_s)} \in \mathbb{R}^{T \times h \times w \times d}$, which preserve fine-grained spatial details within all frames.

Spatiotemporal encoding. Prior to the spatiotemporal attention blocks, we incorporate learnable temporal positional embeddings ($\mathbf{P}_{\text{tem}} \in \mathbb{R}^{T \times d}$) to the spatial features ($\mathcal{F}_{\text{spa}}$) to encode temporal ordering information. We adopt a TimeSformer-like [5] approach to extend the remaining L_{st} attention blocks to spatiotemporal attention, where each block alternately performs temporal attention and spatial attention. For a token at temporal position t and spatial position i , temporal attention can be expressed as:

$$\mathbf{z}_{t,i}^{(l+\frac{1}{2})} = \text{TemporalAttn}(\mathbf{z}_{t,i}^{(l)}, \{\mathbf{z}_{t',i}^{(l)}\}_{t'=1}^T) \quad (7)$$

Here, each token interacts with all tokens at the same spatial position (i), without information exchange across distinct spatial positions. Subsequently, a spatial attention layer is applied, which operates in the same manner as previously mentioned, represented as:

$$\mathbf{z}_{t,i}^{(l+1)} = \text{SpatialAttn}(\mathbf{z}_{t,i}^{(l+\frac{1}{2})}, \{\mathbf{z}_{t,j}^{(l+\frac{1}{2})}\}_{j=1}^N) \quad (8)$$

After L_s spatial and L_{st} spatiotemporal attention blocks, we apply an MAP head (attention pooling) [79] across spatial dimensions to the final features ($\mathbf{z}^{(L_s+L_{st})}$) following SigLIP 2 [61], and obtain the final **semantic features**, denoted as $\mathcal{F}_{\text{sem}} = \text{MAP}(\mathbf{z}^{(L_s+L_{st})}) \in \mathbb{R}^{T \times d}$, that capture global dynamic semantic information.

This hierarchical design allows the model to capture both spatial details and temporal dynamics in soccer videos, yielding unified representations for downstream tasks.

5.3. Supervised Multi-task Pretraining

We pretrain our visual encoder with a multi-task framework, on spatial perception tasks (*e.g.*, athlete detection and identification, pitch registration) and semantic reasoning tasks (*e.g.*, event classification, vision-language alignment), with lightweight output heads ($\Psi_{\text{out}} = \{\Psi_d, \Psi_k, \Psi_l, \Psi_e, \Psi_a\}$), detailed as follows.

Athlete detection and identification. The athlete detection and identification head (Ψ_d) employs a lightweight Deformable DETR decoder-like [83] structure to perform attention between learnable queries and the extracted spatial features ($\mathcal{F}_{\text{spa}}$) to obtain object-level features, which are then passed through three linear layers that predict the athlete bounding box ($\hat{\mathbf{b}}_i$), player role ($\hat{\mathbf{r}}_i$), and jersey number ($\hat{\mathbf{n}}_i$), respectively, expressed as:

$$\{\hat{\mathbf{b}}_i; \hat{\mathbf{r}}_i; \hat{\mathbf{n}}_i\} = \Psi_d(\mathcal{F}_{\text{spa}}) \quad (9)$$

We adopt bounding box regression loss ($\mathcal{L}_{\text{bbox}}$) and classification loss ($\mathcal{L}_{\text{cls}}$) commonly used in detection tasks, while

role classification ($\mathcal{L}_r$) and jersey number recognition ($\mathcal{L}_j$) employ focal loss [34]. The overall objective function for this task is a weighted sum of the above items expressed as:

$$\mathcal{L}_a = \lambda_{\text{cls}}\mathcal{L}_{\text{cls}} + \lambda_{\text{bbox}}\mathcal{L}_{\text{bbox}} + \lambda_r\mathcal{L}_r + \lambda_j\mathcal{L}_j \quad (10)$$

Pitch registration. The pitch registration heads ($\{\Psi_k, \Psi_l\}$) inherit the heatmap-based method in PnLCalib [24], taking the spatial feature ($\mathcal{F}_{\text{spa}}$) as input and using two convolutional networks (CNNs) to predict keypoint and endpoint heatmaps for each frame, expressed as:

$$\{\hat{\mathcal{K}}; \hat{\mathcal{L}}\} = \{\Psi_k(\mathcal{F}_{\text{spa}}); \Psi_l(\mathcal{F}_{\text{spa}})\} \quad (11)$$

Specifically, each head uses convolutional layers containing PixelShuffle [57] to perform progressive upsampling on features, ultimately using MSE losses between predicted keypoints ($\hat{\mathcal{K}}$) and lines ($\hat{\mathcal{L}}$) and their ground truth as the optimization objective, denoted as $\mathcal{L}_k$ and $\mathcal{L}_l$.

Event classification. Following UniSoccer [53], the event classification head (Ψ_e) adopts a two-layer transformer encoder, followed by temporal average pooling and a linear classifier, which predicts soccer event ($\hat{e}$) based on semantic reasoning features ($\mathcal{F}_{\text{sem}}$), expressed as:

$$\hat{e} = \Psi_e(\mathcal{F}_{\text{sem}}) \quad (12)$$

The head is optimized using cross-entropy loss ($\mathcal{L}_e$).

Vision-language alignment. Similar to UniSoccer [53], this alignment head (Ψ_a) first performs temporal average pooling on semantic reasoning features ($\mathcal{F}_{\text{sem}}$) to obtain video representations and computes similarity (s) with text embeddings of commentary ($\mathcal{T}$) extracted by an off-the-shelf SigLIP 2 text encoder (Φ_{text}) [61], expressed as:

$$s = \Psi_a(\mathcal{F}_{\text{sem}}, \Phi_{\text{text}}(\mathcal{T})) \quad (13)$$

The head is optimized using SigLIP loss [80] in a contrastive learning manner, denoted as $\mathcal{L}_{\text{con}}$.

Overall training loss. The final training objective ($\mathcal{L}_{\text{total}}$) combines all task-specific ones with corresponding weights:

$$\mathcal{L}_{\text{total}} = \lambda_a\mathcal{L}_a + \lambda_k\mathcal{L}_k + \lambda_l\mathcal{L}_l + \lambda_e\mathcal{L}_e + \lambda_{\text{con}}\mathcal{L}_{\text{con}} \quad (14)$$

This unified multi-task pretraining enables SoccerMaster to learn complementary spatial and semantic representations.

5.4. Downstream Task Adaptation

After pretraining, our SoccerMaster is seamlessly adapted to various downstream tasks. By introducing task-specific heads, we demonstrate the versatility and generalization ability of our representations across commentary generation, camera calibration, and multiple object tracking.

Commentary generation. Similar to [51, 53], we adopt a Q-Former [31] architecture to perform temporal aggregation on semantic features ($\mathcal{F}_{\text{sem}}$) and employ a linear layer

Table 3. **Performance Comparison on Pretraining Tasks.** We compare SoccerMaster against both general-purpose VFM (SigLIP2-L/16-512 [61] and DINOv3-L/16 [58]) and the soccer-specific MatchVision [53] using frozen encoders with trainable task-specific heads. “SoccerFactory Data” indicates training augmented with automatically generated data from our pipeline alongside existing datasets. Metrics include AP@50 and mAP for detection, jersey number (jn) and role classification accuracy, keypoint/line detection metrics, event classification accuracy, and video-commentary retrieval top-1 accuracy (computed within batches of 48).

Encoder	SoccerFactory Data	Athlete Detection and Identification				Keypoints Detection			Lines Detection			Event Cls.	Alignment
		AP@50	mAP	jn	role	accuracy	precision	recall	accuracy	precision	recall	accuracy	top-1
SigLIP 2	✓	69.1	29.2	<u>78.8</u>	97.5	92.5	80.0	82.0	90.0	80.2	85.9	49.8	3.4
SigLIP 2		<u>72.3</u>	<u>32.0</u>	78.2	97.3	93.7	80.0	<u>83.4</u>	<u>93.6</u>	84.7	<u>92.3</u>	-	-
DINOv3	✓	61.8	22.9	75.6	<u>98.5</u>	92.5	77.2	80.2	91.0	82.0	87.1	51.8	-
DINOv3		<u>70.2</u>	28.0	76.1	98.1	<u>94.0</u>	81.4	<u>83.4</u>	92.8	82.7	92.7	-	-
MatchVision	✓	51.9	17.0	74.9	94.1	92.0	<u>85.2</u>	69.8	90.3	<u>86.6</u>	81.5	<u>65.3</u>	4.0
Ours	✓	91.5	49.5	79.7	99.1	95.3	85.9	84.1	95.7	91.4	92.4	77.2	39.0

to project the pooled tokens into the embedding space of a large language model (Llama-3-8B [23]), which serve as prefix embeddings for auto-regressive text generation.

Camera calibration. Building upon the field keypoints and lines detected by our pitch registration heads ($\{\Psi_k, \Psi_l\}$), we apply the Point and Line (PnL) refinement module [24] to estimate the camera intrinsic and extrinsic parameters.

Multiple object tracking. Following MOTIP [18], we perform tracking by formulating object association as a classification task. We extract object-level features from the DETR decoder of the athlete detection and identification head (Ψ_d). These features are concatenated with learnable ID embeddings from a dictionary to construct historical trajectories, which act as in-context identity prompts. Finally, a transformer-based ID decoder predicts consistent identity labels by attending to historical trajectories.

6. Experiments

This section presents comprehensive evaluations of SoccerMaster across various downstream tasks to validate its effectiveness. We begin with implementation details in Sec. 6.1, followed by the comparisons on downstream tasks in Sec. 6.2, and conduct ablation studies in Sec. 6.3.

6.1. Implementation Details

SoccerMaster model adopts a hierarchical vision transformer architecture, initialized with weights from siglip2-large-patch16-512 [61]. The visual encoder comprises $L_s = 16$ spatial and $L_{st} = 8$ spatiotemporal transformer blocks, with a hidden dimension $d = 1024$. Input videos are uniformly sampled into clips of $T = 30$ frames at 512×512 resolution with a patch size of 16×16 . Further implementation details are provided in the **Supplementary Material**.

6.2. Performance Comparison

Comparison on pretraining tasks. We compare against general-purpose baselines (SigLIP2 [61] and DINOv3 [58]) and the soccer-specific baseline (MatchVision) on the pretraining tasks. Dataset details appear in Tab. 2. For fair comparison, we adopt the following evaluation protocol:

Table 4. **Comparison on Camera Calibration.** Here, JaC_γ , CR, and FS denote the Jaccard index at threshold γ , completeness rate, and final score, respectively. For our model, * denotes zero-shot inference and † indicates fine-tuned training.

Dataset	Method	JaC_γ [%]			CR	FS	Resolution
		5	10	20			
SN22 -test -center	PnLCalib	80.3	91.9	94.2	97.9	78.6	960×540
	PnLCalib	70.2	86.8	91.7	96.4	67.6	512×512
	Ours*	70.5	<u>92.0</u>	94.8	99.4	70.1	512×512
	Ours†	<u>76.9</u>	92.3	94.4	98.6	<u>75.8</u>	512×512
SN23 -test	PnLCalib	76.7	87.2	90.1	79.5	60.9	960×540
	PnLCalib	66.9	82.4	86.7	78.4	51.8	512×512
	Ours*	59.5	78.1	81.6	75.2	44.8	512×512
	Ours†	<u>71.1</u>	<u>85.4</u>	<u>87.8</u>	<u>79.1</u>	<u>56.2</u>	512×512

for baseline models, we freeze their pretrained encoders and train only lightweight task-specific heads on our pretraining dataset; While for **SoccerMaster**, as it has been pretrained on these tasks, we directly evaluate it. This setup assesses the representation quality of each encoder by maintaining consistent task-specific head architectures.

As shown in Tab. 3: (i) SoccerMaster significantly outperforms baselines, with **+17.5%** mAP in athlete detection and **+11.9%** accuracy in event classification; (ii) SoccerMaster exhibits remarkable performance in the vision-language alignment, achieving a **39.0%** top-1 accuracy that surpasses all baselines. General-purpose models like SigLIP 2 suffer from severe domain gaps (only 3.4% accuracy), while DINOv3 lacks a corresponding text encoder for alignment evaluation; and (iii) although tailored for soccer data, MatchVision still lacks comprehensive task coverage. Despite being competitive in keypoint detection, line detection, and event classification, it struggles with both athlete detection and identification and vision-language alignment, likely due to an over-reliance on joint vision-language optimization that neglects fine-grained spatial perception and misalignment with the pre-trained SigLIP 2 text encoder.

Comparison on camera calibration. We evaluate camera calibration performance against the state-of-the-art method, namely PnLCalib [24], on the SoccerNet-22 and SoccerNet-23 benchmarks. Following standard protocols [20, 24, 39], we assess performance using the Jaccard index (JaC_γ),

Table 5. Ablation Study on the Impact of Automatically Generated Spatial Annotations.

SoccerFactory Data	Athlete Detection and Identification				Keypoints Detection			Lines Detection			Event Classification	Alignment
	AP@50	mAP	jn	role	accuracy	precision	recall	accuracy	precision	recall	accuracy	top-1
✓	77.7	30.2	75.0	97.3	92.9	89.6	70.5	93.4	93.9	86.0	71.6	35.0
	82.0	37.5	76.5	98.1	93.1	90.1	70.7	93.7	96.1	85.2	70.5	36.8

completion rate (CR), and the final score (FS), detailed in the **Supp. Mat.**. As shown in Tab. 4, PnlCalib is optimized for a higher resolution of 960×540 , achieving state-of-the-art performance in that setting (78.6 FS on SN22-test-center and 60.9 FS on SN23-test). However, since SoccerMaster operates at a 512×512 resolution, we evaluate all methods at this resolution to ensure a fair comparison.

The results reveal two observations: (i) on zero-shot evaluation, SoccerMaster already outperforms PnlCalib on SN22-test-center (70.1 vs. 67.6 FS). On the more challenging SN23-test, which contains diverse non-main camera views, our zero-shot model achieves 44.8 FS compared to PnlCalib’s 51.8 FS; (ii) once fine-tuned, SoccerMaster establishes a new standard at this resolution, surpassing PnlCalib by **+8.2** FS on SN22 (75.8 vs. 67.6) and **+4.4** FS on SN23 (56.2 vs. 51.8). This confirms SoccerMaster learns strong, transferable representations for field geometry.

Comparison on multiple object tracking. As shown in Tab. 6, we evaluate our method on the SoccerNet-tracking benchmark [9] using standard MOT metrics (HOTA [37], DetA [37], AssA [37], MOTA [4] and IDF1 [55]). Unlike the tracking-by-detection paradigm relying on complex multi-stage components, SoccerMaster employs a simple end-to-end pipeline while achieving competitive results. Overall, our model demonstrates competitive results across all key metrics: HOTA (59.1), DetA (65.2), AssA (53.9), MOTA (81.6), and IDF1 (74.6). These results highlight the strong transferability of our representations, which can be effectively adapted to tracking tasks through lightweight fine-tuning, significantly simplifying the inference pipeline.

Table 6. Comparison on Multiple Object Tracking. Notably, our model is the only one employing an end-to-end setting.

Method	HOTA	DetA	AssA	MOTA	IDF1
YOLOv8+OC-SORT [6]	54.6	<u>63.5</u>	47.1	<u>76.2</u>	62.5
YOLOv8+StrongSort++ [14]	56.2	62.9	50.3	75.0	66.5
YOLOv8+PRTreID [41]	59.8	61.1	58.6	73.1	<u>74.5</u>
Ours+MOTIP [18]	<u>59.1</u>	65.2	<u>53.9</u>	81.6	74.6

Table 7. Comparison on Commentary Generation.

Visual Encoder	BLEU@1	BLEU@4	METEOR	ROUGE-L	CIDEr
SigLIP 2	28.9	6.9	24.7	25.4	28.3
MatchVision	<u>30.9</u>	<u>8.7</u>	26.9	27.6	<u>35.7</u>
Ours	31.3	8.9	<u>26.2</u>	<u>26.6</u>	38.6

Comparison on commentary generation. We evaluate this task on SN-Caption-test-align [51], comprising 49 matches with accurate temporal alignment and manual an-

notations. For fair comparison, all models are fine-tuned on the MatchTime [51] dataset. Tab. 7 shows results using metrics of language generation, including BLEU [46], METEOR [3], ROUGE-L [33], and CIDEr [64]. SoccerMaster outperforms MatchVision on BLEU@1 (**+0.4**) and CIDEr (**+2.9**), achieving the best overall performance. While METEOR and ROUGE-L scores are marginally lower than MatchVision’s, our model surpasses the general-purpose SigLIP 2 [61] baseline. The superior CIDEr score confirms that our multi-task pretraining yields rich semantic representations for language generation.

6.3. Ablation Study

Tab. 5 evaluates the effectiveness of our automatically generated spatial annotations. For computational efficiency, we use a compact SoccerMaster variant: 224×224 input, 16×16 patch size, and $L_s = 8, L_{st} = 4$ encoding layers. SoccerFactory-generated data yields substantial gains in spatial perception tasks, especially in athlete detection with **+4.3** AP@50 (82.0 vs. 77.7) and **+7.3** mAP (37.5 vs. 30.2). Other pretraining tasks, including pitch registration, event classification, and vision-language alignment, maintain stable or slightly improved performance. This confirms that generated annotations provide a reliable and scalable training source. More ablation studies and qualitative results are in the **Supplementary Material**.

7. Conclusion

This paper aims to advance soccer visual understanding by proposing a unified framework. Concretely, we present **SoccerMaster**, the first soccer-specific vision foundation model capable of handling comprehensive tasks spanning fine-grained spatial perception and high-level semantic reasoning through supervised multi-task pretraining. To facilitate large-scale pretraining, we introduce **SoccerFactory**, an automated data curation pipeline that generates scalable spatial annotations from broadcast videos. Integrating these annotations with multiple existing soccer datasets, we construct a comprehensive pretraining dataset. Extensive evaluations demonstrate that SoccerMaster serves as an effective soccer-specific vision foundation model, where simple fine-tuning yields state-of-the-art performance across diverse downstream tasks. We believe this work establishes a new paradigm for soccer visual understanding by offering a unified, adaptable framework that reduces the engineering overhead of developing task-specific systems and provides a solid foundation for future advances in soccer AI.

Acknowledgments

Weidi would like to acknowledge the funding from Scientific Research Innovation Capability Support Project for Young Faculty (ZY-GXQNJSKYCXNLZCXM-I22).

References

- [1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 3
- [2] Bavesh Balaji, Jerrin Bright, Harish Prakash, Yuhao Chen, David A Clausi, and John Zelek. Jersey number recognition using keyframe identification from low-resolution broadcast videos. In *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, 2023. 1, 2
- [3] 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*, 2005. 8
- [4] Keni Bernardin and Rainer Stiefelwagen. Evaluating multiple object tracking performance: the clear mot metrics. *EURASIP Journal on Image and Video Processing*, 2008. 8
- [5] 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*, 2021. 5, 6
- [6] Jinkun Cao, Jiangmiao Pang, Xinshuo Weng, Rawal Khrodkar, and Kris Kitani. Observation-centric sort: Rethinking sort for robust multi-object tracking. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 8
- [7] 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*, 2021. 1
- [8] 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*, 2022. 1
- [9] 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*, 2022. 1, 2, 8
- [10] Anthony Cioppa, Silvio Giancola, Vladimir Somers, Victor Joos, Floriane Magera, Jan Held, Seyed Abolfazl Ghasemzadeh, Xin Zhou, Karolina Seweryn, Mateusz Kowalczyk, et al. Soccernet 2024 challenges results. *arXiv preprint arXiv:2409.10587*, 2024. 2
- [11] Yutao Cui, Chenkai Zeng, Xiaoyu Zhao, Yichun Yang, Gangshan Wu, and Limin Wang. Sportsmot: A large multi-object tracking dataset in multiple sports scenes. In *Proceedings of the International Conference on Computer Vision*, 2023. 2
- [12] 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*, 2021. 1, 2, 3, 4
- [13] 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. 2, 5
- [14] Yunhao Du, Zhicheng Zhao, Yang Song, Yanyun Zhao, Fei Su, Tao Gong, and Hongying Meng. Strongsort: Make deepsort great again. *IEEE Transactions on Multimedia*, 2023. 3, 8
- [15] Nikolay S Falaleev and Ruilong Chen. Enhancing soccer camera calibration through keypoint exploitation. In *ACM Multimedia Workshops*, 2024. 2, 3
- [16] Roya Firoozi, Johnathan Tucker, Stephen Tian, Anirudha Majumdar, Jiankai Sun, Weiyu Liu, Yuke Zhu, Shuran Song, Ashish Kapoor, Karol Hausman, et al. Foundation models in robotics: Applications, challenges, and the future. *The International Journal of Robotics Research*, 2025. 2
- [17] Haoxiang Gao, Zhongruo Wang, Yaqian Li, Kaiwen Long, Ming Yang, and Yiqing Shen. A survey for foundation models in autonomous driving. *arXiv preprint arXiv:2402.01105*, 2024. 2
- [18] Ruopeng Gao, Ji Qi, and Limin Wang. Multiple object tracking as id prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025. 7, 8
- [19] 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*, 2018. 1, 2
- [20] Silvio Giancola, Anthony Cioppa, Adrien Deliege, Floriane Magera, Vladimir Somers, Le Kang, Xin Zhou, Olivier Barnich, Christophe De Vleeschouwer, Alexandre Alahi, et al. Soccernet 2022 challenges results. In *Proceedings of the 5th International ACM Workshop on Multimedia Content Analysis in Sports*, 2022. 2, 7
- [21] Silvio Giancola, Anthony Cioppa, et al. Soccernet 2025 challenges results. *arXiv preprint arXiv:2508.19182*, 2025. 2, 4
- [22] Vladimir Golovkin, Nikolay Nemtsev, Vasyli Shandyba, Oleg Udin, Nikita Kasatkin, Pavel Kononov, Anton Afanasiev, Sergey Ulasen, and Andrei Boiarov. From broadcast to minimap: Achieving state-of-the-art soccernet game state reconstruction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2025. 4
- [23] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha

- Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024. 7
- [24] Marc Gutiérrez-Pérez and Antonio Agudo. Pnlib: Sports field registration via points and lines optimization. *arXiv preprint arXiv:2404.08401*, 2024. 2, 3, 6, 7
- [25] 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*, 2023. 1, 2
- [26] Jan Held, Hani Itani, Anthony Cioppa, Silvio Giancola, Bernard Ghanem, and Marc Van Droogenbroeck. X-vars: Introducing explainability in football refereeing with multi-modal large language models. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 1, 2
- [27] Mostafa S. Ibrahim, Srikanth Muralidharan, Zhiwei Deng, Arash Vahdat, and Greg Mori. A hierarchical deep temporal model for group activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 2
- [28] Maria Koshkina and James H. Elder. A general framework for jersey number recognition in sports video. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 1, 3, 4
- [29] 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 preprint arXiv:2401.01505*, 2024. 2
- [30] 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*, 2022. 2
- [31] 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*, 2023. 2, 6
- [32] Yixuan Li, Lei Chen, Runyu He, Zhenzhi Wang, Gangshan Wu, and Limin Wang. Multisports: A multi-person video dataset of spatio-temporally localized sports actions. In *Proceedings of the International Conference on Computer Vision*, 2021. 2
- [33] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, 2004. 8
- [34] Tsung-Yi Lin, Priya Goyal, Ross Girshick, Kaiming He, and Piotr Dollár. Focal loss for dense object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2017. 6
- [35] Yikun Liu, Yuan Liu, Shangzhe Di, Haicheng Wang, Zhongyin Zhao, Le Tian, Xiao Zhou, Jie Zhou, Jiangchao Yao, Yanfeng Wang, et al. Versavit: Enhancing mllm vision backbones via task-guided optimization. *arXiv preprint arXiv:2602.09934*, 2026. 2
- [36] Zhaoyu Liu, Kan Jiang, Murong Ma, Zhe Hou, Yun Lin, and Jin Song Dong. F³set: Towards analyzing fast, frequent, and fine-grained events from videos. In *Proceedings of the International Conference on Learning Representations*, 2025. 2
- [37] Jonathon Luiten, Aljosa Osep, Patrick Dendorfer, Philip Torr, Andreas Geiger, Laura Leal-Taixé, and Bastian Leibe. Hota: A higher order metric for evaluating multi-object tracking. *International Journal of Computer Vision*, 2021. 8
- [38] Chunyan Ma, Jizhuang Fan, Jing-Yue Yao, and Tao Zhang. Npu rgbd dataset and a feature-enhanced lstm-dgc method for action recognition of basketball players+. *Applied Sciences*, 2021. 2
- [39] Floriane Magera, Thomas Hoyoux, Olivier Barnich, and Marc Van Droogenbroeck. A universal protocol to benchmark camera calibration for sports. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2024. 1, 2, 7
- [40] Floriane Magera, Thomas Hoyoux, Olivier Barnich, and Marc Van Droogenbroeck. Broadtrack: Broadcast camera tracking for soccer. In *Winter Conference on Applications of Computer Vision*, 2025. 1
- [41] 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 *Proceedings of the 6th International Workshop on Multimedia Content Analysis in Sports*, 2023. 1, 2, 3, 8
- [42] Weibo Mao, Chenxin Xu, Qi Zhu, Siheng Chen, and Yanfeng Wang. Leapfrog diffusion model for stochastic trajectory prediction. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2
- [43] 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*, 2023. 1, 2
- [44] Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, et al. Dinov2: Learning robust visual features without supervision. *Transactions on Machine Learning Research*, 2024. 2
- [45] Yulu Pan, Ce Zhang, and Gedas Bertasius. Basket: A large-scale video dataset for fine-grained skill estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025. 2
- [46] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Association for Computational Linguistics*, 2002. 8
- [47] AJ Piergiovanni and Michael S. Ryoo. Fine-grained activity recognition in baseball videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2018. 2
- [48] 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*, 2023. 1, 2
- [49] 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*, 2020. 2
- [50] 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
- [51] 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, 4, 6, 8
- [52] Jiayuan Rao, Zifeng Li, Haoning Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Multi-agent system for comprehensive soccer understanding. In *ACM Multimedia*, 2025. 1, 2
- [53] Jiayuan Rao, Haoning Wu, Hao Jiang, Ya Zhang, Yanfeng Wang, and Weidi Xie. Towards universal soccer video understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025. 1, 2, 3, 4, 6, 7
- [54] Nikhila Ravi, Valentin Gabeur, Yuan-Ting Hu, Ronghang Hu, Chaitanya Ryali, Tengyu Ma, Haitham Khedr, Roman Rädle, Chloe Rolland, Laura Gustafson, et al. Sam 2: Segment anything in images and videos. In *Proceedings of the International Conference on Learning Representations*, 2025. 2, 3
- [55] Ergys Ristani, Francesco Solera, Roger Zou, Rita Cucchiara, and Carlo Tomasi. Performance measures and a data set for multi-target, multi-camera tracking. In *Proceedings of the European Conference on Computer Vision*, 2016. 8
- [56] 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*, 2020. 2
- [57] Wenzhe Shi, Jose Caballero, Ferenc Huszár, Johannes Totz, Andrew P Aitken, Rob Bishop, Daniel Rueckert, and Zehan Wang. Real-time single image and video super-resolution using an efficient sub-pixel convolutional neural network. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2016. 6
- [58] Oriane Siméoni, Huy V Vo, Maximilian Seitzer, Federico Baldassarre, Maxime Oquab, Cijo Jose, Vasil Khalidov, Marc Szafraniec, Seungeun Yi, Michaël Ramamonjisoa, et al. Dinov3. *arXiv preprint arXiv:2508.10104*, 2025. 2, 7
- [59] 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*, 2024. 1, 2, 3, 4
- [60] 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*, 2017. 2
- [61] Michael Tschannen, Alexey Gritsenko, Xiao Wang, Muhammad Ferjad Naeem, Ibrahim Alabdulmohsin, Nikhil Parthasarathy, Talfan Evans, Lucas Beyer, Ye Xia, Basil Mustafa, et al. Siglip 2: Multilingual vision-language encoders with improved semantic understanding, localization, and dense features. *arXiv preprint arXiv:2502.14786*, 2025. 2, 3, 6, 7, 8
- [62] Renaud Vandeghen, Anthony Cioppa, and Marc Van Droogenbroeck. Semi-supervised training to improve player and ball detection in soccer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, 2022. 1, 2
- [63] Rejin Varghese and M Sambath. Yolov8: A novel object detection algorithm with enhanced performance and robustness. In *International Conference on Advances in Data Engineering and Intelligent Computing Systems*, 2024. 3
- [64] 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*, 2015. 8
- [65] Jianyuan Wang, Minghao Chen, Nikita Karaev, Andrea Vedaldi, Christian Rupprecht, and David Novotny. Vggv: Visual geometry grounded transformer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025. 2
- [66] Chaoyi Wu, Xiaoman Zhang, Ya Zhang, Hui Hui, Yanfeng Wang, and Weidi Xie. Towards generalist foundation model for radiology by leveraging web-scale 2d&3d medical data. *Nature Communications*, 2025. 2
- [67] 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
- [68] 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.
- [69] 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.
- [70] Zeyu Xi, Haoying Sun, Yaofei Wu, Junchi Yan, Haoran Zhang, Lifang Wu, Liang Wang, and Changwen Chen. Player-centric multimodal prompt generation for large language model based identity-aware basketball video captioning. In *Proceedings of the International Conference on Computer Vision*, 2025. 2
- [71] 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

- [72] 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](#)
- [73] Huangbiao Xu, Xiao Ke, Huanqi Wu, Rui Xu, Yuezhou Li, and Wenzhong Guo. Language-guided audio-visual learning for long-term sports assessment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2025. [2](#)
- [74] 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*, 2022. [2](#)
- [75] Yibin Yan, Jilan Xu, Shangzhe Di, Yikun Liu, Yudi Shi, Qirui Chen, Zeqian Li, Yifei Huang, and Weidi Xie. Learning streaming video representation via multitask training. In *Proceedings of the International Conference on Computer Vision*, 2025. [2](#)
- [76] Yibin Yan, Jilan Xu, Shangzhe Di, Haoning Wu, and Weidi Xie. Omnistream: Mastering perception, reconstruction and action in continuous streams. *arXiv preprint arXiv:2603.12265*, 2026. [2](#)
- [77] Lihe Yang, Bingyi Kang, Zilong Huang, Zhen Zhao, Xiaogang Xu, Jiashi Feng, and Hengshuang Zhao. Depth anything v2. In *Conference on Neural Information Processing Systems*, 2024. [2](#)
- [78] Ling You, Wenxuan Huang, Xinni Xie, Xiangyi Wei, Bangyan Li, Shaohui Lin, Yang Li, and Changbo Wang. Timesoccer: An end-to-end multimodal large language model for soccer commentary generation. In *ACM Multimedia*, 2025. [2](#)
- [79] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2022. [2](#), [6](#)
- [80] 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*, 2023. [6](#)
- [81] Chenhao Zheng, Jieyu Zhang, Mohammadreza Salehi, Ziqi Gao, Vishnu Iyengar, Norimasa Kobori, Quan Kong, and Ranjay Krishna. One trajectory, one token: Grounded video tokenization via panoptic sub-object trajectory. In *Proceedings of the International Conference on Computer Vision*, 2025. [2](#)
- [82] Xiao Zhou, Luoyi Sun, Dexuan He, Wenbin Guan, Ruifen Wang, Lifeng Wang, Xin Sun, Kun Sun, Ya Zhang, Yanfeng Wang, et al. A knowledge-enhanced pathology vision-language foundation model for cancer diagnosis. In *Proceedings of the European Conference on Computer Vision*, 2024. [2](#)
- [83] Xizhou Zhu, Weijie Su, Lewei Lu, Bin Li, Xiaogang Wang, and Jifeng Dai. Deformable detr: Deformable transformers for end-to-end object detection. *Proceedings of the International Conference on Learning Representations*, 2020. [6](#)