SoccerMaster: A Vision Foundation Model for Soccer Understanding

Supplementary Material

Contents

A. Pretraining Dataset Construction and Integration

To facilitate comprehensive multi-task pretraining, we construct a unified, large-scale dataset by integrating the automatically curated data from our pipeline, **SoccerFactory**, with multiple existing soccer video datasets [4, 14–16]. This section details the datasets used for each pretraining task, their preprocessing strategies, and integration into our unified training framework. Specifically, we first detail the data automatically curated by SoccerFactory in Sec. A.1, then introduce the integration of spatial and semantic tasks in Sec. A.2 and Sec. A.3, along with the statistics of the integrated dataset in Sec. A.4.

A.1. Details of SoccerFactory-Generated Data

Our proposed automated curation pipeline, **SoccerFactory**, generates high-quality, frame-level annotations for spatial perception tasks on soccer broadcast videos. Specifically, each video clip is accompanied by a `.json` file containing dense per-frame annotations of athletes, field keypoints, and lines, structured as follows:

Frame-level annotations. Each frame in the video is associated with holistic spatial annotations, including athlete detection, pitch registration, and role classification information:

```
{
  "athletes": [...],                         # List of athlete annotations
  "keypoints": [...],                        # Pitch keypoints
  "lines": [...],                            # Pitch line segments
  "K": [...],                                # Camera intrinsic matrix
  "Rt": [...],                               # Camera extrinsic matrix
  "valid_cam_params": True                   # Calibration validity flag
}
```

Athlete annotations. For each athlete detected in a video frame, we provide the attributes describing this player, including bounding box coordinates, track ID, jersey number, legibility score, role, and team affiliation, for example:

```
{
  "bbox_ltwh": [1116.5, 679.5, 50.8, 98.2],  # Bounding box (left, top, width, height)
  "track_id": 4,                             # Tracklet unique identity
  "jersey_number": "10",                     # Jersey number (tracklet-level)
  "legibility_score": 0.67,                  # Jersey number legibility score
  "role": "player",                          # Player role
  "team": "right"                            # Team affiliation
}
```

Field keypoint annotations. We provide 2D keypoints detected by the field keypoint detector, indexed according to the semantic definitions in PnlCalib [7]. Here is an example:

```
{
  2: {"x": 984.0, "y": 348.0, "p": 0.800},   # Coordinates and confidence
  32: {"x": 984.0, "y": 460.0, "p": 0.846},
  ...
}
```

Field line annotations. We provide detected field line segments as sequences of 2D points in normalized image coordinates. Each line is identified by its semantic label following SoccerNet-v3 [2]. An example is as follows:

```
{
  "Circle central": [                        # Sequences of points
    {"x": 0.513, "y": 0.426},
    {"x": 0.388, "y": 0.441},
    {"x": 0.329, "y": 0.470}, ...
  ],
  "Middle line": [
    {"x": 0.513, "y": 0.322},
    {"x": 0.513, "y": 0.426},
    {"x": 0.515, "y": 0.485}, ...
  ],
  ...
}
```

A.2. Spatial Perception Tasks

During pretraining, the spatial perception module addresses two primary tasks: (i) athlete detection and identification, and (ii) pitch registration. These tasks leverage annotated data from two sources: automatically generated data from our pipeline and existing data from SoccerNet-GSR [16].

A.2.1. Data Sources Introduction

SoccerFactory-generated data are produced by our automated data curation pipeline, with annotation format details provided in Sec. A.1. We extract main-camera clips from 500 matches in the SoccerNetv2 [4] dataset by utilizing the ground truth from the SoccerNet Camera Shot Segmentation task [4], specifically selecting segments labeled as "main camera center" to exclude replays, close-ups, and alternative angles. To align with SoccerNet-GSR [16], these extracted clips are further partitioned into segments with a maximum duration of 30 seconds.

SoccerNet-GSR [16] comprises 200 uncut, 30-second broadcast video clips captured by a single moving camera, providing realistic and challenging scenarios for spatial understanding tasks. It features extensive annotations, including over 9.37 million line points for pitch localization and camera calibration, as well as over 2.36 million athlete positions on the pitch labels with their roles, teams, and jersey numbers.

A.2.2. Datasets Integration Details

Data preprocessing and integration. For both pipeline-generated data and SoccerNet-GSR, we retain a unified subset of annotations per frame used in pretraining. Specifically, each frame contains the following information:

```
{
  "athletes": [...],                          # List of athlete annotations
  "keypoints": [...],                         # Pitch keypoints
  "lines": [...]                              # Pitch line segments
}
```

For each athlete detection, we maintain the following structured data, for example:

```
{
  "bbox_ltwh": [1116.5, 679.5, 50.8, 98.2],  # Bounding box
  "track_id": 4,                              # Track identity
  "jersey_number": "10",                      # Jersey number (bbox-level)
  "role": "player"                            # Player role
}
```

Jersey number filtering. Since SoccerNet-GSR provides tracklet-level jersey number annotations, we employ a legibility classifier [9] to compute frame-level legibility scores for each detection. For both datasets, jersey numbers with legibility scores below 0.5 are set to `null`. This aligns with our athlete detection task definition, which specifies that jersey numbers should be marked as `null` when they are not clearly visible due to occlusion or viewpoint variations.

Pitch keypoints and lines generation. For pitch registration annotations, we adopt different strategies for the two datasets based on their characteristics. For SoccerNet-GSR, we directly utilize the manually annotated pitch keypoints and line segments provided in the dataset. For data produced by our pipeline, we leverage the estimated camera parameters (intrinsic matrix $\mathbf{K}$ and extrinsic matrix $[\mathbf{R}|\mathbf{t}]$) to project a standard soccer pitch model containing line segments onto the 2D image plane, and then extract the visible portion within the frame boundaries to serve as ground truth annotations. This projection-based approach for our curated data provides geometrically consistent and precise annotations at scale, while the manual annotations from SoccerNet-GSR offer high-quality ground truth for validation and fine-grained supervision.

A.3. Semantic Reasoning Tasks

In our multi-task pretraining, the semantic reasoning tasks focus on event classification and vision-language alignment. We utilize data from three datasets: SoccerNet-v2 [4], MatchTime [14], and SoccerReplay-1988 [15], as detailed below.

A.3.1. Data Sources Introduction

SoccerNet-v2. This dataset contains over 110k event labels across 500 matches, originally categorized into 17 distinct classes. Following UniSoccer [15], we systematically remap these labels into 24 standardized event categories. For example, "Direct free-kick" and "Indirect free-kick" are merged into a single "Free Kick" category, while "Penalty" outcomes are distinguished as "Penalty" and "Penalty Missed."

Table 1. **Statistics of the Integrated Pretraining Dataset.** An overview of the train/validation/test sample counts for all datasets employed in multi-task pretraining. Here, data generated by SoccerFactory is exclusively used for training. Meanwhile, SoccerNet-GSR additionally includes samples from 36 sequences for the SoccerNet challenge evaluation.

Dataset	Train	Valid	Test	Tasks
Spatial Perception Tasks (Dense Sampling, 25 FPS)				
SoccerNet-GSR [16]	1,425	1,450	1,225	Athlete Detection Pitch Registration
SoccerFactory-generated	94,628	–	–	Athlete Detection Pitch Registration
Semantic Reasoning Tasks (Sparse Sampling, 1 FPS)				
SoccerNet-v2 [4]	54,448	17,491	18,641	Event Classification
MatchTime [14]	24,027	3,144	3,256	Event Classification Vision-Language Alignment
SoccerReplay-1988 [15]	104,080	17,892	17,402	Event Classification Vision-Language Alignment

MatchTime. The MatchTime dataset [14] consists of 471 matches, featuring high-quality aligned video-commentary pairs, which is the temporally aligned version of SoccerNet-Caption [13]. To leverage this dataset for both event classification and vision-language alignment, we adopt the prompt-based summarization approach from UniSoccer [15] to extract event categories from the commentary text, thereby associating each video clip with a corresponding event label.

SoccerReplay-1988. This large-scale dataset provides 1,988 matches with both event labels and temporally aligned commentaries, serving as the primary resource for joint training in event classification and vision-language alignment.

A.3.2. Datasets Integration Details

We integrate all samples from these three datasets into a unified training corpus. While all three datasets provide event labels for event classification training, only MatchTime and SoccerReplay-1988 include text commentaries. Consequently, during training, the event classification loss is computed on all samples, whereas the vision-language alignment loss is calculated exclusively on samples with available commentaries. The final training utilizes 30-second soccer video clips alongside their corresponding event labels (spanning 24 classes) and commentaries when applicable. The train/valid/test split settings remain consistent with the original dataset configurations.

A.4. Data Statistics

To accommodate the different nature of spatial perception and semantic reasoning tasks, we employ tailored sampling strategies for each category. For **spatial perception tasks**, we sample dense frames from video clips at 25 FPS, where each training sample comprises 30 consecutive frames. Conversely, for **semantic reasoning tasks**, we adopt a sparse sampling strategy at 1 FPS; here, each video clip constitutes a single sample centered around the specific event timestamp or commentary moment. Tab. 1 presents the detailed statistics of data samples across all datasets within our multi-task pretraining framework. The resulting training resource, termed **SoccerFactory**, aggregates these diverse data to facilitate comprehensive pretraining.

B. More Implementation Details

B.1. Implementation details of SoccerMaster

SoccerMaster is initialized from SigLIP 2-large-patch16-512 [17], with zero-initialized temporal attention layers and temporal positional embeddings to enable temporal modeling while preserving the pretrained spatial features. The model is optimized using the AdamW [10] optimizer with differentiated learning rates across model components, as detailed in Tab. 2. All parameters except temporal positional embeddings are regularized with a weight decay of 1.0×10^{-4}. These specific learning rates account for the distinct initialization states and functional roles of each module. We train for a total of 20 epochs, with the first epoch using linear warm-up and subsequent epochs following a CosineAnnealingLR [8] schedule.

To stabilize multi-task training, we apply gradient normalization and clipping independently to the backbone and each task-specific head. By clipping the gradient norm to a maximum of 0.1 for all components, we prevent any single task from dominating the optimization process. The multi-task learning objective integrates various loss components with carefully

Table 2. **Learning Rates for Model Components.** Here, temporal-related components* include temporal attention layers and temporal positional embeddings, which are zero-initialized.

Component	Backbone (Temporal-agnostic)	Backbone (Temporal-related)*	Athlete Detection Head	Keypoint Detection Head	Line Detection Head	Event Classification Head	Vision-Language Alignment
Learning Rate	5.0×10^{-5}	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-4}	1.0×10^{-3}	1.0×10^{-4}	1.0×10^{-4}

Table 3. **Loss Weights for Multi-task Training.**

Loss Component	**Weight**	**Scope**
Athlete Detection Sub-losses		
Classification Loss	$\lambda_{\text{cls}} = 2.0$	Within detection head
Bounding Box Regression Loss	$\lambda_{\text{bbox}} = 5.0$	Within detection head
Role Classification Loss	$\lambda_{\text{r}} = 2.0$	Within detection head
Jersey Number Recognition Loss	$\lambda_{\text{j}} = 1.0$	Within detection head
Task-level Losses		
Athlete Detection	$\lambda_{\text{a}} = 1.0$	Task-level
Keypoint Detection	$\lambda_{\text{k}} = 1.0$	Task-level
Line Detection	$\lambda_{\text{l}} = 1.0$	Task-level
Event Classification	$\lambda_{\text{e}} = 1.0$	Task-level
Vision-Language Alignment	$\lambda_{\text{con}} = 4.0$	Task-level

Table 4. **Data Augmentation Configuration.** Here, ± denotes symmetric ranges (*e.g.*, ±10° represents [-10°, 10°]).

Augmentation Type	**Probability**	**Parameters**
Geometric Transformations		
Random Affine	0.5	Rotation: ±10°, Translation: ±10%, Scale: [0.9, 1.1], Shear: ±5°
Random Perspective	0.5	Distortion scale: 0.3
Random Crop	1.0	Size ratio: [0.6, 1.0]
Horizontal Flip	0.5	–
Photometric Transformations		
Color Jitter	1.0	Brightness: 0.4, Contrast: 0.4 Saturation: 0.4, Hue: 0.2
Gaussian Noise	0.3	Standard deviation: 0.02
Gaussian Blur	0.2	Kernel size: [3, 7], Sigma: [0.1, 2.0]

tuned weights, as shown in Tab. 3. These weights serve to balance optimization signals across tasks characterized by varying sample sizes and loss magnitudes.

As specified in Tab. 4, we employ a data augmentation strategy incorporating both geometric and photometric transformations to improve model robustness to varying viewpoints, lighting conditions, and image quality. All pretraining experiments are conducted on $16\times$ NVIDIA H800 GPUs, utilizing a global batch size of 16 for spatial perception tasks and 32 for semantic reasoning tasks. The complete multi-task pretraining requires approximately 9 days in full precision (*float32*).

B.2. Implementation details of SoccerFactory

The field keypoint and line detectors, along with the StrongSORT [5] and PRTReID [12] components, are off-the-shelf models utilized in accordance with the SoccerNet-GSR [16] baseline; specifically, the former use the *SV_kp* and *SV_lines* weights from [7], while the latter adopt weights and hyperparameters fine-tuned on the SoccerNet-tracking dataset [3] from [12]. For detection, we fine-tune YOLOv8x6 [18] on the SoccerNet-GSR training set. For attribute recognition, we leverage Qwen2.5-VL [1]: the 7B variant is used for athlete role classification, and the 72B variant is employed for jersey number recognition to ensure superior performance, with post-hoc filtering performed by an off-the-shelf legibility classifier [9] using a confidence threshold of 0.5.

Table 5. **Ablation Study on the Impact of Multi-task Pretraining.**

Multi-task	Athlete Detection				Keypoints Detection			Lines Detection			Event Classification	Alignment
	AP@50	mAP	jn	role	accuracy	precision	recall	accuracy	precision	recall	Accuracy	Top-1
	88.3	**43.8**	**78.5**	**98.7**	93.0	90.0	**70.9**	**93.9**	**96.2**	**85.5**	69.1	32.6
✓	82.0	37.5	76.5	98.1	**93.1**	**90.1**	70.7	93.7	96.1	85.2	**70.5**	**36.8**

C. Details of Camera Calibration Metrics

In our camera calibration evaluation, we strictly adhere to the evaluation protocol in previous works [6, 7, 11], and report the Jaccard index (JaC_γ) at thresholds of 5, 10, and 20 pixels, completion rate (CR) and final score (FS). Specifically, the Jaccard index (JaC_γ) measures calibration accuracy based on reprojection error; it is calculated as TP/(TP+FN+FP), where a pitch line is deemed correctly detected only if all its points exhibit reprojection errors below the threshold γ. The completion rate (CR) quantifies the proportion of images for which the method successfully produces camera parameters, while the final score, defined as $\mathrm{FS} = \mathrm{CR} \times \mathrm{JaC}_5$, serves as the primary evaluation criterion.

D. Ablation on Multi-task Pretraining

To investigate the impact of multi-task pretraining on model performance across different tasks, we conduct an ablation study by comparing two training strategies: (i) single-task training, where each task is learned independently, and (ii) multi-task training, where all tasks are jointly optimized. To ensure computational efficiency, both experiments are conducted using our compact SoccerMaster variant.

According to the results presented in Tab. 5, we draw the following observations: (i) For keypoint detection, line detection, and event classification, both strategies yield comparable performance (within a 1% margin). (ii) However, multi-task pretraining substantially improves vision-language alignment (top-1: $32.6\% \rightarrow 36.8\%$), indicating that joint optimization provides complementary supervisory signals for semantic understanding. (iii) Conversely, athlete detection suffers noticeable degradation (AP@50: $88.3\% \rightarrow 82.0\%$, mAP: $43.8\% \rightarrow 37.5\%$), which we attribute to the limited capacity of the compact variant struggling with large-scale multi-task optimization.

To validate this hypothesis, we train our full-scale SoccerMaster exclusively on athlete detection and identification, and compare it with the multi-task variant shown in Tab. 3 in the main paper. The single-task model yields performance levels comparable to our multi-task model (-0.3% AP@50, -0.5% mAP), while showing marginal improvements in jersey number accuracy ($+0.8\%$) and role accuracy ($+0.1\%$). These negligible fluctuations confirm that, given sufficient model capacity, multi-task pretraining can achieve comparable performance on individual tasks while simultaneously learning robust representations for diverse downstream applications.

E. Qualitative Results

E.1. Qualitative Results of SoccerFactory

Fig. 1 presents a visualization example of SoccerFactory on the SoccerNet-GSR [16] test set, comparing predictions (left) against ground truth annotations (right). Each detected athlete is annotated with five attributes: ID (tracklet identity), R (role), L (legibility score), JN (jersey number), and T (team affiliation). Regarding field registration, detected keypoints and field lines are highlighted in yellow and red, respectively, while lines projected from the canonical pitch coordinate system via estimated camera parameters are shown in blue. In the ground truth images, manually labeled field lines are displayed in red.

These results demonstrate that SoccerFactory achieves robust performance across diverse scenarios. The system maintains high accuracy in role classification, jersey number recognition, and team affiliation, even under challenging conditions such as crowded scenes and varying camera viewpoints. Notably, SoccerFactory exhibits strong temporal consistency in tracking. Specifically, the first and third rows of Fig. 1 correspond to the first and last frames of the same video clip, respectively. Despite significant camera motion and player movement across frames, SoccerFactory successfully maintains consistent identity assignments for the tracked athletes, highlighting the effectiveness of our tracking and post-processing modules.

Fig. 2 presents athlete positions mapped from image coordinates to standardized pitch coordinates using the estimated camera parameters. The visualization displays a bird's-eye view of the pitch, where the bottom-center of each athlete's bounding box is projected onto the 2D pitch coordinate system to represent their ground-plane position. Athletes are color-coded by role: referees (orange, labeled "RE"), the left team (red), and the right team (blue). Non-referee athletes are further distinguished by their tracklet identities. Note that while specific identity values may differ between predictions and ground truth, the primary evaluation criterion is temporal consistency, *i.e.*, ensuring the same athlete retains the same identity across frames. The close alignment between our predictions and ground truth annotations demonstrates the pipeline's capability to accurately estimate camera parameters and maintain consistent athlete tracking across frames. This precise localization in pitch coordinates can potentially facilitate downstream applications such as tactical analysis [19, 20], showcasing the practical value of our automatic data curation pipeline beyond basic detection and tracking tasks.

E.2. Qualitative Results of SoccerMaster

We present qualitative results of SoccerMaster on a video clip in Fig. 3, demonstrating its comprehensive understanding capabilities across multiple soccer understanding tasks. The results for multiple object tracking and commentary generation are obtained with task-specific heads and downstream task adaptation. The prediction conventions follow the same format as described in Fig. 1, with generated commentary displayed at the bottom of each frame.

The model exhibits strong performance across athlete detection, pitch registration, multiple object tracking, event classification, and commentary generation. Notably, even in crowded scenes, the system maintains robust detection and tracking capabilities, successfully identifying individual athletes and preserving temporal identity consistency across frames. Furthermore, despite minor detection errors in keypoint and line predictions in some frames, the camera calibration remains accurate, as evidenced by the close alignment between the projected pitch lines (blue) and the actual pitch lines and goal frames visible in the RGB images. This demonstrates the model's robustness to local prediction errors through the geometric refinement process. Overall, these results validate SoccerMaster's capability as a unified foundation model to handle diverse soccer understanding tasks within a single framework.

F. Limitations & Future Works

F.1. Limitations

While SoccerMaster demonstrates strong performance across diverse soccer understanding tasks, several limitations remain to be addressed in future work.

Jersey number recognition. As illustrated in Fig. 3, our model occasionally produces incorrect jersey number predictions. We formulate jersey number recognition as a simple 101-class classification problem (digits 0-99 and `null`), where the `null` class dominates the training data due to frequent occlusions and non-frontal viewpoints in broadcast footage. This severe class imbalance, combined with limited sample diversity for visible jersey numbers, hinders accurate jersey number recognition in challenging scenarios. Furthermore, unlike traditional methods that apply dedicated recognition models to cropped bounding boxes, our approach performs recognition jointly with detection in a single forward pass. While our unified approach offers greater efficiency, the two-stage approach benefits from processing high-resolution crops focused on jersey details, potentially achieving higher accuracy at the cost of increased computational overhead.

Goalkeeper classification. Another notable issue is the misclassifications of goalkeepers as players, which can be observed in several frames of Fig. 3. This problem stems from two primary factors: (i) goalkeepers constitute only a small fraction of athletes in the training data, leading to insufficient samples for learning discriminative features; and (ii) goalkeeper uniforms exhibit high variance in color across matches and leagues. Distinguishing them from players based solely on individual appearance is difficult without relational reasoning that compares uniform colors across all detected athletes.

Limited scope of spatial perception. Our current framework focuses exclusively on athlete detection and pitch registration, without considering ball detection and tracking.

F.2. Future Works

To address the aforementioned limitations and further enhance SoccerMaster's capabilities, our future work will focus on refining the data curation pipeline. While our current approach is effective, we aim to generate higher-quality annotations at scale by integrating more sophisticated tracking algorithms and enforcing stricter temporal consistency constraints. Additionally, we plan to incorporate auxiliary modalities, such as audio commentary and player statistics, to improve annotation accuracy. We anticipate that scaling up the training dataset through this enhanced data curation pipeline will yield substantial performance gains across all tasks.

Figure 1. **Qualitative Results of SoccerFactory.** Comparison between our predictions (left) and ground truth annotations (right) on the SoccerNet-GSR test set. Our pipeline demonstrates robust performance across diverse scenarios.

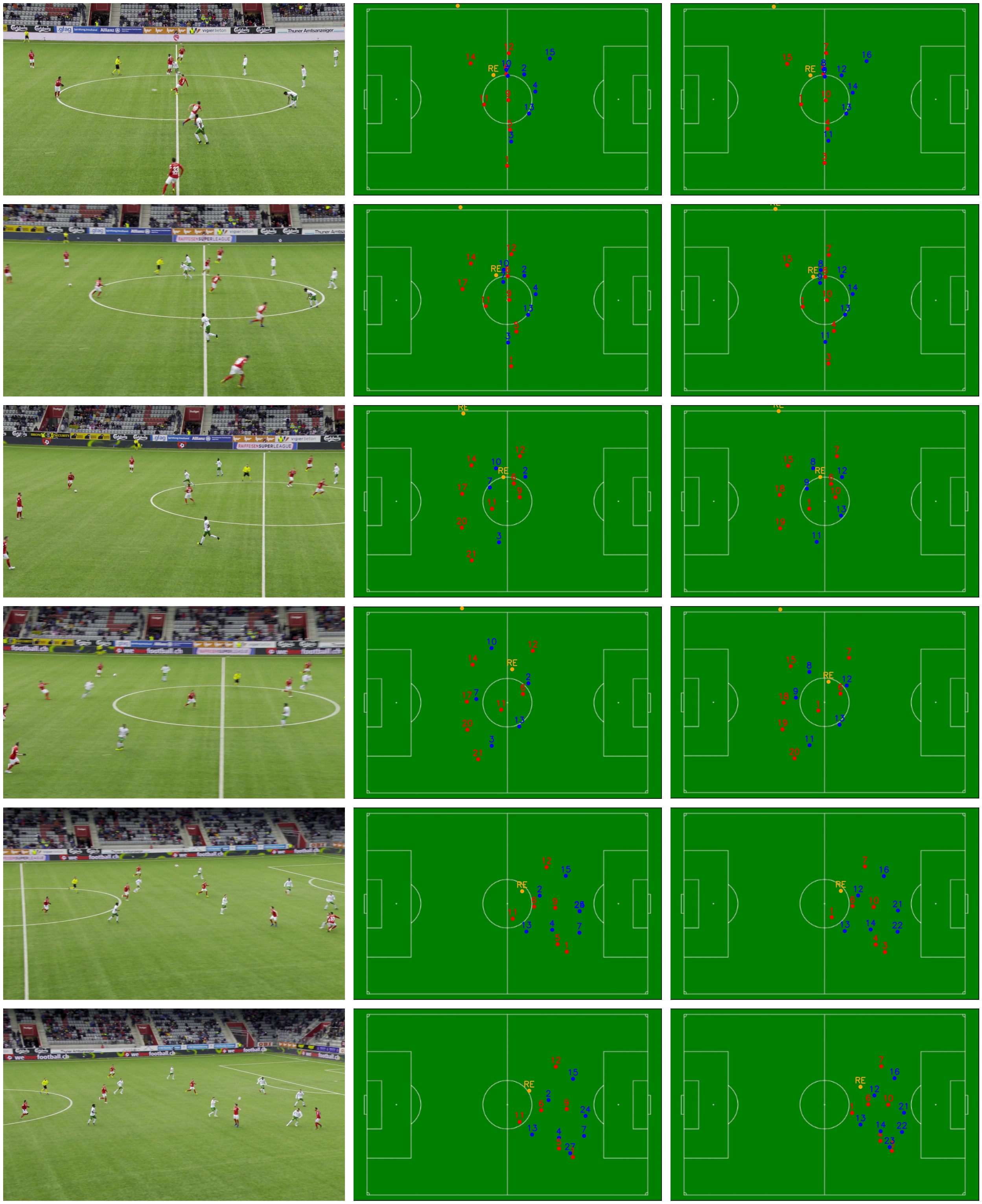
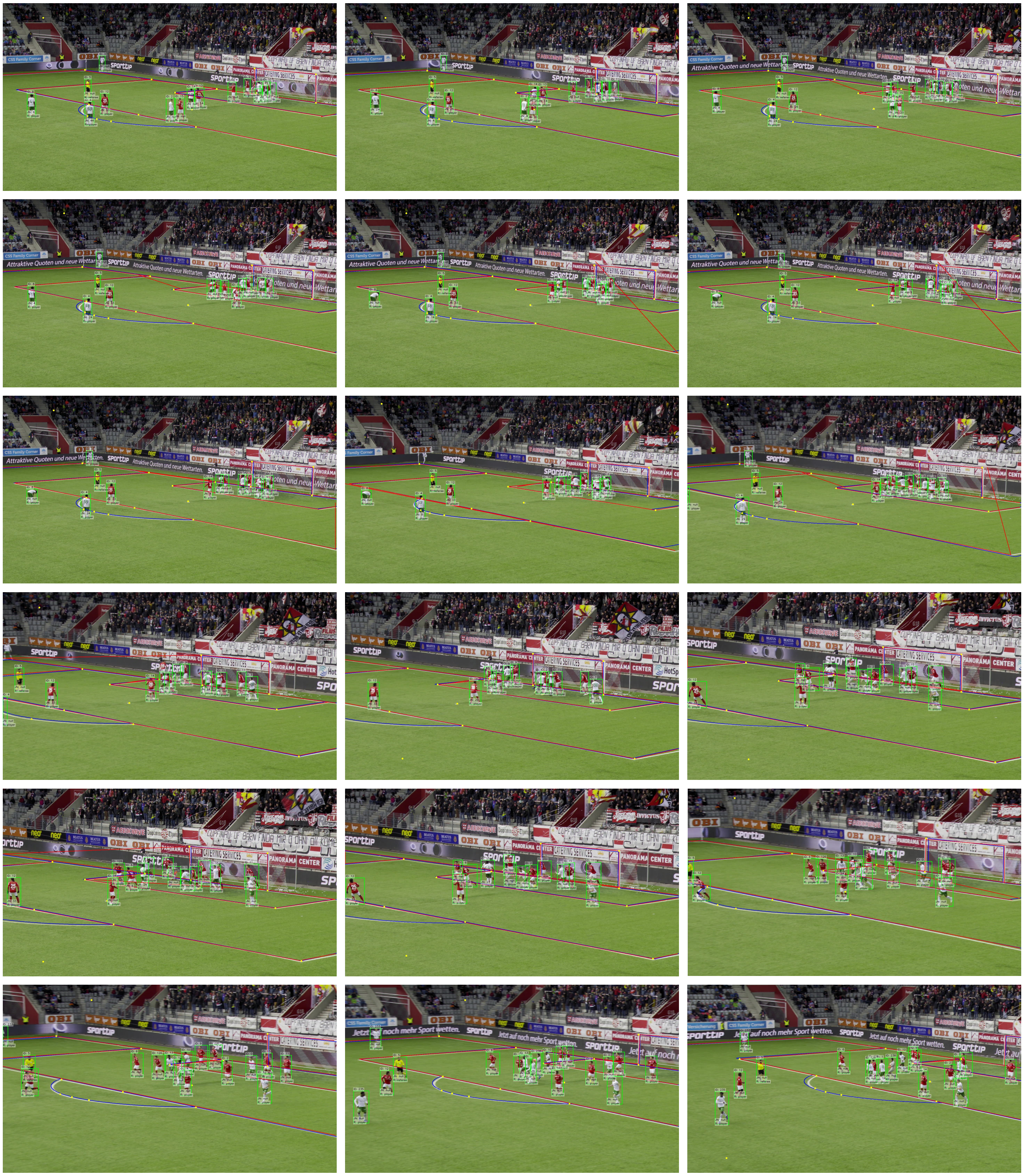


Figure 2. **Top-view Pitch Visualization of SoccerFactory Results.** Athlete positions are mapped to standardized pitch coordinates via estimated camera parameters. Each row is organized as: input image (left), our predictions (middle), and ground truth annotations (right). Athletes are color-coded by role: referees (orange, labeled "RE"), left team (red), and right team (blue). Non-referee athletes are labeled with arbitrary **tracklet identities**, which maintain temporal consistency within each tracking sequence. Notably, the number of tracklet identities in the predictions is not directly comparable to those in the ground truth; the emphasis is on consistent identity assignment across frames.



Event: *Corner*
Commentary: *[PLAYER] ([TEAM]) takes the corner kick and sends the ball into the penalty area, but the opposition's defense is ready and clears the danger.*

Figure 3. **Qualitative Results of SoccerMaster.** SoccerMaster can simultaneously execute multiple soccer understanding tasks on a video clip, including athlete detection, pitch registration, multiple object tracking, event classification, and commentary generation. Frames are arranged in temporal order from left to right and top to bottom.

References

[1] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 5

[2] 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. 2

[3] 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. 5

[4] 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. 2, 3, 4

[5] 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. 5

[6] Silvio Giancola, Anthony Cioppa, Adrien Deliège, 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. 6

[7] Marc Gutiérrez-Pérez and Antonio Agudo. Pnlcalib: Sports field registration via points and lines optimization. *arXiv preprint arXiv:2404.08401*, 2024. 2, 5, 6

[8] Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 4

[9] 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. 3, 5

[10] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *Proceedings of the International Conference on Learning Representations*, 2019. 4

[11] 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. 6

[12] 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. 5

[13] 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. 4

[14] 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. 2, 3, 4

[15] 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. 3, 4

[16] 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. 2, 3, 4, 5, 6

[17] 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. 4

[18] 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. 5

[19] 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*, 2024. 7

[20] Siyao Zhao, Hao Ma, Zhiqiang Pu, Jingjing Huang, Yi Pan, and Zhi Ming. Taceleven: generative tactic discovery for football open play. *arXiv preprint arXiv:2511.13326*, 2025. 7