# FOOTBOTS: A TRANSFORMER-BASED ARCHITECTURE FOR MOTION PREDICTION IN SOCCER

*Guillem Capellera*[1,2] *Luis Ferraz*[1] *Antonio Rubio*[1] *Antonio Agudo*[2] *Francesc Moreno-Noguer*[2]

[1] Kognia Sports Intelligence [2] Institut de Robòtica i Informàtica Industrial, CSIC-UPC

## ABSTRACT

Motion prediction in soccer involves capturing complex dynamics from player and ball interactions. We present FootBots, an encoder-decoder transformer-based architecture addressing motion prediction and conditioned motion prediction through equivariance properties. FootBots captures temporal and social dynamics using set attention blocks and multi-attention block decoder. Our evaluation utilizes two datasets: a real soccer dataset and a tailored synthetic one. Insights from the synthetic dataset highlight the effectiveness of FootBots' social attention mechanism and the significance of conditioned motion prediction. Empirical results on real soccer data demonstrate that FootBots outperforms baselines in motion prediction and excels in conditioned tasks, such as predicting the players based on the ball position, predicting the offensive (defensive) team based on the ball and the defensive (offensive) team, and predicting the ball position based on all players. Our evaluation connects quantitative and qualitative findings. https://youtu.be/9kaEkfzG3L8

***Index Terms*—** Motion prediction, Signal forecasting, Transformer, Trajectory understanding, Soccer.

## 1. INTRODUCTION

Multi-agent Motion Prediction (MP) holds critical importance across various domains, encompassing financial economics [1], human pose estimation [2, 3, 4, 5], pedestrian behavior analysis [6, 7, 8, 9, 10, 11], and sports analytics [12, 13, 14, 15, 16, 17]. This task involves forecasting future positions and motions of multiple agents in a shared environment. In multi-agent sports like soccer, accurate motion prediction deepens insights into player behavior, team dynamics, and on-field decision-making processes.

The intricacies of soccer, characterized by swift changes in player positions, rapid ball movements, and intricate team coordination, underscore the need for advanced models surpassing conventional non-social-aware motion prediction. In image processing and computer vision, these challenges crucially apply to enhancing player tracking and re-identification, where precise position forecasts contribute to performance analysis, team strategies, and overall gameplay understanding. These forecasts can also be used to simulate team strategies and obtain advanced metrics [18]. The dynamics of soccer, coupled with player interactions, drive the Conditioned Motion Prediction (CMP) task, adressing scenarios like predicting player trajectories based on ball position [13, 17]. Figure 1 illustrates the MP task and four different CMP tasks in detail that we consider in this paper.

Given the dynamic nature of soccer, a robust model must exhibit permutation equivariance [13, 14], adapting to varying player compositions and interactions. Transformer-based architectures [19], known for their handling of varying-length sequences and permutation equivariance, have gained traction in motion prediction tasks [10, 11], including sports [17].

This research introduces a comprehensive approach to soccer MP and CMP, employing a transformer-based model. Our model captures soccer's intricate properties, leveraging permutation equivariance and historical data to predict ball and player 2D trajectories. Evaluation against baselines showcases social awareness in soccer. The study introduces a synthetic dataset tailored for this research, and utilizes a real one from LaLiga 2022-2023.

## 2. RELATED WORK

The evolution of multi-agent motion prediction originates from human pose techniques, initially utilizing Recurrent Neural Networks (RNN) [2, 3] to capture temporal dynamics. This evolution extended to pedestrian motion, fusing RNN with social pooling [6, 7] to capture social interactions. However, [20] introduced an RNN baseline void of social encoding, yet surpassing social pooling methods.

Recognizing the significance of social interactions led to the adoption of Graph Neural Networks (GNN) in pedestrian modeling [8, 9], coupled with recurrent techniques. Additionally, siMLPe [5] demonstrates the effectiveness of a Multi-Layer Perceptron (MLP) architecture in capturing temporal and spatial dynamics for human motion prediction. Transformer-based architectures also made substantial contributions, showcasing their ability to encode temporal and social dynamics in human pose estimation [4] and in forecasting pedestrian and vehicle trajectories [10, 11, 21].

This work has been supported by the project MoHuCo PID2020-120049RB-I00 funded by MCIN/AEI/10.13039/501100011033 and by the Government of Catalonia under 2023 DI 00058.

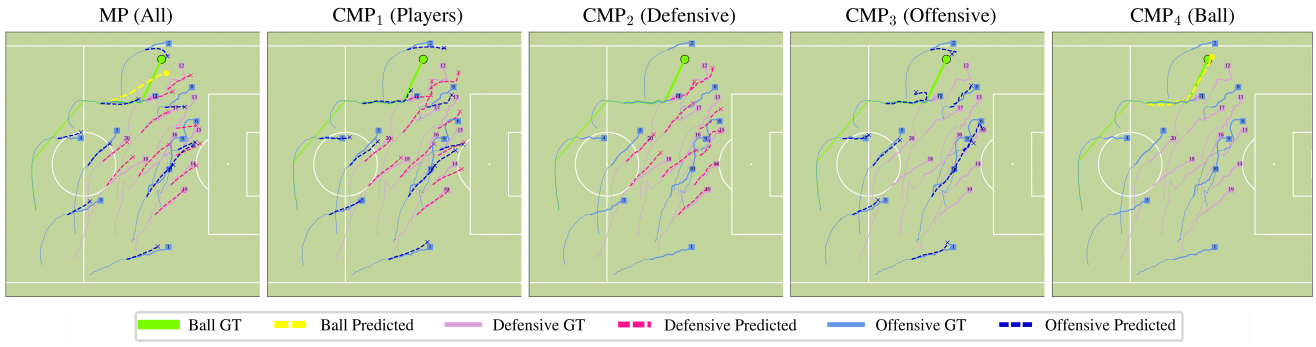


**Fig. 1**. **Motion prediction in soccer.** The method predicts both player and ball motions from partial 2D trajectories under specified conditions. In the figure, squares represent the end positions of ground truth offensive and defensive team players, crosses denote their predicted positions, and circles indicate the final ball ones. Five different tasks (MP, $\text{CMP}_{1-4}$) for the same test sequence are displayed, everyone of them is tailored to predict specific subsets of agents, as specified in parentheses.

In sports, initial methods focused on generating long-term basketball trajectories using Variational Autoencoders (VAE) [22] and Variational Recurrent Neural Networks (VRNN) [23, 24]. These variational methods were later outperformed by a multi-modal RNN-based architecture [15]. To leverage permutation equivariance without the necessity of agent ordering, subsequent research integrated VRNN with GNN for generating multi-agent trajectories in basketball and soccer [13, 25, 16]. Addressing concerns associated with accumulated errors in recurrent methods, [14] combined Graph Attention Networks (GAT) with temporal convolutional networks in soccer. Transformer-based models have also found applications in the sports domain, demonstrating superior performance compared to graph-recurrent-based approaches [17, 26] in NBA trajectories. Nevertheless, simultaneously conducting attention in both temporal and social dimensions still incurs a notable computational cost.

Our method introduces a tailored transformer encoder-decoder for soccer, adeptly adapting to the sport's intricacies involving a higher number of agents compared to basketball. To enhance computational efficiency, the model is optimized by sequentially decoupling temporal and social attentions. We leverage permutation equivariance alongside the agents' ordering. Moreover, we showcase the effectiveness of our approach in addressing soccer's CMP task using a tailored synthetic dataset and a real one, skillfully capturing intricate agent interactions.

## 3. OUR METHOD

### 3.1. Problem statement

Consider a set of $M \in \mathbb{N}$ agent measures ($M-1$ players and a ball in our context), $X = \{\mathbf{x}^1, \dots, \mathbf{x}^M\}$ where every measure contains $k$ elements. The measures evolve over a time horizon of $t+T \in \mathbb{N}$ where $t$ and $T$ are positive integers. Particularly, $t$ represents the observations, while $T$ covers predictions spanning approximately 4 seconds [12, 13, 14]. On the one hand, we can define both prior $\mathcal{X}_{0:t}$ and posterior $\mathcal{X}_{t+1:t+T}$ sequences where specific $\mathbf{x}_t^m$ are collected to define our MP problem as:

$$f(\mathcal{X}_{0:t}) = \mathcal{X}_{t+1:t+T}, \tag{1}$$

where $f$ represents a function to infer the posterior data from the prior.

On the other hand, we also consider a CMP problem. In particular, CMP predicts the positions of some specific agents ($P$), given the positions of other ones ($C$) in a scenario. Let $\mathcal{X}^C_{0:t+T}$ represent the complete sequence of observations for the $C$ agents, and let $\mathcal{X}^P_{0:t}$ and $\mathcal{X}^P_{t+1:T}$ denote the prior and posterior states of the agents to be predicted, respectively. Then we can define the model $f_c$ to sort out the CMP as:

$$f_c(\mathcal{X}^C_{0:t+T}, \mathcal{X}^P_{0:t}) = \mathcal{X}^P_{t+1:t+T}. \tag{2}$$

The challenge in MP is to forecast trajectories of all $M$ agents (players and ball). In contrast, CMP encompasses various tasks within soccer that we introduce next:

**$\text{CMP}_1$**: Predicting players' positions with the ball as a conditioning agent.

**$\text{CMP}_2$**: Predicting defensive team players' positions using all other agents as conditioning ones.

**$\text{CMP}_3$**: Predicting offensive team players' positions using all other agents as conditioning ones.

**$\text{CMP}_4$**: Predicting the ball position with players as conditioning agents.

### 3.2. Attention mechanisms

Attention mechanisms are effective at capturing relationships in sequences or sets. Given $n$ queries $\mathbf{Q}$ of dimension $d_k$, $n_v$ keys $\mathbf{K}$ of dimension $d_k$, and $n_v$ values $\mathbf{V}$ of dimension $d_v$, the attention mechanism computes weighted value sums using compatibility between queries and keys as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_k}}\right)\mathbf{V}, \tag{3}$$

with $\mathbf{Q} \in \mathbb{R}^{n \times d_k}$, $\mathbf{K} \in \mathbb{R}^{n_v \times d_k}$, $\mathbf{V} \in \mathbb{R}^{n_v \times d_v}$. In practice, the attention mechanism is often extended with multiple at-

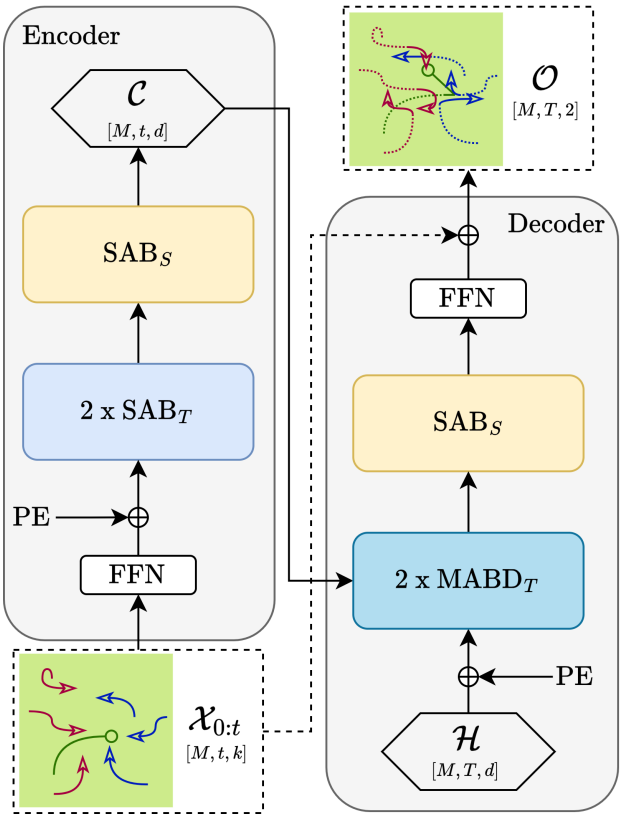


**Fig. 2. FootBots architecture in soccer.** FootBots exploits an encoder-decoder structure with sequential temporal and social attention mechanisms. It incorporates Set Attention Blocks to encode temporal SAB$_T$ and social SAB$_S$ dynamics represented in the context $\mathcal{C}$. The Multi-Attention Block Decoder in the temporal axis (MABD$_T$) and SAB$_S$ in the decoder generate the predicted trajectories. FootBots is capable of solving both MP and CMP tasks in soccer, with an input of the decoder $\mathcal{H}$ varying depending on the task.

tention heads, also called Multi-Head Attention (MHA), originally introduced in Transformer architecture [19], allowing the model to capture different relations in the data.

The MHA operation was extended to work on sets by defining a Set Attention Block (SAB) [27], an adaptation of the encoder block of the Transformer that lacks the positional encoding. The MHA itself provides the property of permutation equivariance, making the SAB a permutation-equivariant operation. Finally, the original Transformer [19] builds the output using its decoder, also called Multi-Attention Block Decoder (MABD) [11], which utilizes cross-attention to take into account the encoder output.

In motion prediction, attention mechanisms can capture temporal dynamics and social interactions among agents. Temporal attention focuses on sequence dynamics, assigning varying weights to observations for accurate future motion prediction. Social attention complements that by considering interactions among agents, capturing spatial relationships, and accounting for collaborative behaviors.

### 3.3. FootBots

In designing FootBots, we leverage SAB and MABD blocks drawing inspiration from [11, 10]. FootBots utilizes an encoder-decoder structure with sequential temporal and social attention mechanisms, capturing player-ball dynamics over time. Figure 2 illustrates its key components.

The encoder of FootBots operates by handling input observations denoted as $\mathcal{X}_{0:t}$ through a Feed-Forward Network (FFN), which is supplemented with a positional encoder (PE) to ensure the temporal ordering of the data. To capture the dynamics in both time and social interactions, we utilize SAB. More specifically, SAB$_T$ for temporal encoding dynamics and SAB$_S$ for social encoding. Those blocks are responsible for capturing interactions among the players and the ball during the prior state, resulting in the generation of a representation tensor called context, denoted as $\mathcal{C}$.

Given the prior sequence of sets $\mathcal{X}_{0:t} = (X_0, \ldots, X_t)$ with dimensions $[M, t, k]$, we define the encoder operations as follows:

$$\mathcal{C} = \text{SAB}_S\left(\text{SAB}_T(\text{SAB}_T(\text{PE} + \text{FFN}(\mathcal{X}_{0:t})))\right), \tag{4}$$

where $\mathcal{C}$ has the dimension $[M, t, d]$, and $d$ is the chosen dimension for the embeddings.

In the decoder, FootBots generates predicted trajectories $\mathcal{O}$ to approximate $\mathcal{X}_{t+1:t+T}$. To this end, it employs a MABD in the temporal axis (MABD$_T$) and a SAB$_S$ in the social one. MABD$_T$ takes into account the output of the encoder $\mathcal{C}$ and the input of the decoder $\mathcal{H}$, which depends on the task at hand: 1) in MP, $\mathcal{H}$ relies on the last $T$ time steps of $\mathcal{C}$; 2) in CMP, $\mathcal{H}$ incorporates the observations of the conditioning agents during the prediction interval $\mathcal{X}^C_{t+1:t+T}$, along with the last $T$ time steps of $\mathcal{C}$ for the agents of interest ($P$), $\mathcal{C}^P_{t-T:t}$.

Given $T \leq t$, where $T$ represents the desired frames to predict, we outline the decoder operations in the following equation:

$$\mathcal{O} = \mathcal{X}_t + \text{FFN}\left(\text{SAB}_S\left(\text{MABD}_T(\text{MABD}_T(\text{PE} + \mathcal{H}, \mathcal{C}), \mathcal{C})\right)\right), \tag{5}$$

where $\mathcal{X}_t$ is the last set of observations of the prior and $\mathcal{H}$ is one input of the MABD$_T$ operation whose definition depends on the task: $\mathcal{H} = \mathcal{C}_{t-T:t}$ for MP; and $\mathcal{H} = \text{FFN}(\mathcal{X}^C_{t+1:t+T}) \cup \mathcal{C}^P_{t-T:t}$ for CMP. It is worth noting that these relations justify our model's constraint that $T \leq t$.

The dimension of $\mathcal{O}$ is $[M, T, 2]$. A skip connection enables residual learning, particularly benefiting early frame precision. In MP task, FootBots maintains agent permutation equivariance, while in CMP task, it demonstrates partial equivariance for both conditioning agents ($C$) and agents of interest ($P$), preserving this property within their respective subsets.

For completeness, we also propose a non-social variant of our approach denoted by FootBots NS. In this case, our method substitutes social SAB$_S$ attention with temporal SAB$_T$ one, omitting social interactions to showcase their importance.

### 3.4. Loss

The loss function utilized is an Average Displacement Error (ADE), a widely used loss metric in trajectory prediction tasks. It computes the average $l_2$-norm between the predicted trajectories and the ground truth (GT) ones as:

$$\text{ADE} = \frac{1}{MT} \sum_{m=1}^{M} \sum_{j=t+1}^{t+T} \left\| \hat{\mathbf{x}}_j^m - \mathbf{x}_j^m \right\|_2 , \tag{6}$$

where $\hat{\mathbf{x}}_j^m$ and $\mathbf{x}_j^m$ represent the predicted and its corresponding GT position of the $m$-th agent at $j$-th time step.

## 4. EXPERIMENTAL EVALUATION

In this section, we present our experimental results on motion prediction and provide a comparison with respect to competing approaches. For quantitative evaluation, we consider a subset of agents $\hat{M}$ by using four types of metrics. First, we consider the $\text{ADE}_{\hat{M}}$ metric in Eq. (6) for just the set $\hat{M}$.

Second, we propose a Final Displacement Error (FDE) that measures the final deviation between the prediction and the corresponding ground truth location as:

$$\text{FDE}_{\hat{M}} = \frac{1}{|\hat{M}|} \sum_{m \in \hat{M}} \left\| \mathbf{x}_{t+T}^m - \hat{\mathbf{x}}_{t+T}^m \right\|_2 .$$

For completeness, we also propose to compute the Maximum Error (MaxErr) as:

$$\text{MaxErr}_{\hat{M}} = \frac{1}{|\hat{M}|} \sum_{m \in \hat{M}} \max_{j \in \{t+1, \ldots, t+T\}} \left\| \mathbf{x}_j^m - \hat{\mathbf{x}}_j^m \right\|_2 ,$$

and the missing rate (MR) to show the percentage of predictions having an $l_2$-norm greater than 1 meter as:

$$\text{MR}_{\hat{M}} = \frac{1}{|\hat{M}|T} \sum_{m \in \hat{M}} \sum_{j=t+1}^{t+T} \mathbb{I} \left[ \left\| \mathbf{x}_j^m - \hat{\mathbf{x}}_j^m \right\|_2 > 1 \right] ,$$

with $\mathbb{I}(\cdot)$ an indicator function.

### 4.1. Datasets

In this paper, we propose to validate our model on synthetic and real datasets. Next, we provide the most important details for each of them.

**Synthetic dataset**: Created to aid investigation and model development for both MP and CMP tasks, this dataset contains 10,000 training and 1,500 validation sequences. Each sequence spans 20 time steps: 10 for prior ($t$) and the rest for target ($T$). Five agents, including four players and one ball, compose each sequence. The ball follows a linear trajectory initially, but can randomly change direction at a chosen time step, followed by another linear path. Noise is introduced for trajectory variability, causing slight deviations from linearity. Player behaviors encompass remaining stationary with noise (S), linear trajectories with noise (L), and non-linear paths influenced by the ball's position as an attractor (A). The number of players for each behavior is randomly determined, all within a bounded square region $[-15, 15] \times [-15, 15]$ resembling meters in real world. In this dataset, each agent's observation is limited to its $xy$ location, leading to $k = 2$ according to the problem formulation.

**Real dataset:** This dataset comprises actual data from 283 matches of LaLiga's 2022-2023 season, capturing agent motions using advanced computer vision techniques. Each match is divided into sequences representing 9.6 seconds, down-sampled to 6.25 frames per second. Each sequence, consisting of 60 frames, is divided into prior states (35 frames or 5.6 seconds) and target ones (25 frames or 4 seconds). Only trajectories of all 20 field players (excluding goalkeepers) are considered, and agent order is standardized. The dataset is split into 243 matches (82,954 sequences) for training, 20 matches (6,258 sequences) for testing, and 20 matches (7,500 sequences) for validation. Trajectories are normalized to fit within $[-1, 1] \times [-1, 1]$ by dividing by the largest pitch dimension, and spatial realignment ensures the possession team's rightward motion on the pitch. When using FootBots and its non-social variant FootBots NS, each agent's observation includes its 2D position and an associated integer indicating its role: ball, defensive team player, or offensive team player. Therefore, in this case we consider $k = 3$ elements. In contrast, when using other baseline models, input is confined to the 2D position, leading to $k = 2$.

### 4.2. Synthetic evaluation

This initial scenario aims to emphasize the motivation and importance of effectively solving the CMP tasks in soccer. By analyzing the synthetic dataset, we can evaluate the significance of the social $\text{SAB}_S$ and its ability to address the specific $\text{CMP}_1$ task.

The outcomes of our proposed methods on the synthetic dataset are detailed in Table 1, highlighting the performance metrics for both MP and $\text{CMP}_1$ tasks. In terms of $\text{ADE}_P$, FootBots exhibit a slight advantage over FootBots NS when addressing the MP task. This divergence can be attributed to FootBots' capacity to anticipate the motions of ball-attracted players (A) by leveraging the anticipated ball position, facilitated by the social $\text{SAB}_S$. However, it is important to note that despite these enhancements, the persistence of high $\text{ADE}_{\text{ball}}$ values implies that predictions for type A players may be subject to error propagation. Shifting to the $\text{CMP}_1$ task, a notable improvement in predictions is observed.

Nevertheless, for a deeper evaluation, qualitative analysis of example sequences is crucial. Figure 3 provides further insight, showcasing two instances from the validation set, each featuring distinct complexities in ball trajectory. These instances enable differentiation between FootBots NS and Foot-

| Model | Task | Predict ($P$) | ADE$_P$(m) ↓ | ADE$_{\text{ball}}$(m) ↓ |
|---|---|---|---|---|
| FootBots NS | MP | Players+Ball | 0.50 | 1.83 |
| FootBots | MP | Players+Ball | 0.44 | 1.72 |
| FootBots | CMP$_1$ | Players | **0.10** | - |

**Table 1. Evaluating our architecture on synthetic data.**

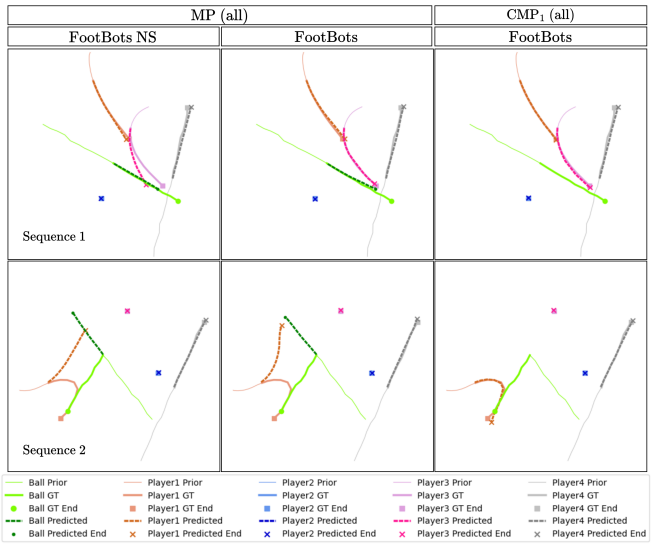


**Fig. 3. Two examples from the synthetic dataset.** The examples serve to visually compare the performance of FootBots NS and FootBots solving MP task, and FootBots solving CMP$_1$. The predictions for different player types (S, L, and A) are evaluated, emphasizing the impact of incorporating social attention and the ball as the conditioning agent.

Bots in the MP task, as well as between FootBots in MP and FootBots in CMP$_1$. All three models adeptly predict static (type S) and linear (type L) player positions with commendable accuracy. However, FootBots NS faces challenges in accurately predicting type A player actions due to its reliance on extrapolating their past trajectories without considering ball-related factors. In the context of the MP task with FootBots, a noteworthy correlation emerges between predictions for type A players and the quality of ball prediction. Consequently, the precision of type A predictions is significantly reliant on accurate ball prediction. This is exemplified in sequence 1, where the ball trajectory retains linearity throughout the prediction time-frame, leading to almost precise predictions. Conversely, in sequence 2, deviations in ball prediction propagate errors to the predictions of type A players. The robustness of our model's capacity to effectively address CMP$_1$ task (conditioned on ball information), is emphasized in the concluding segments of our study. Across the two scenarios presented, the model consistently achieves nearly precise predictions within the CMP$_1$ context.

### 4.3. Real evaluation

In the second evaluation scenario, we analyze the efficacy of FootBots in addressing the MP task by employing a real dataset. We conduct a comparative assessment with various baseline methods to gauge its performance. Additionally, utilizing the same real dataset, we evaluate FootBots' performance in the CMP tasks. The considered baselines, including the already described **FootBots NS**, are outlined as follows:

**Velocity**: We employ velocity extrapolation as a preliminary benchmark, projecting agent predictions linearly based on observed velocity.

**RNN**: We implement an RNN with LSTM cells, using an encoder for input representation and an MLP decoder for prediction, a method proven effective in prior work [20].

**siMLPe**: Adapted from human pose prediction, siMLPe treats each joint as an agent, utilizing an MLP-based model with layer normalization and transposition operations for spatial-temporal dynamics using a fixed sequence length [5].

**baller2vec++**: Adapted from the basketball context, this approach conducts attention simultaneously in both temporal and social dimensions by modelling the attention mask [26].

It is important to note that Velocity, RNN [20], and FootBots NS operate independently for each agent, making the order of the input irrelevant. However, siMLPe [5] is a role-based model and not equivariant. In baller2vec++, they demonstrate minimal result variation when changing the ordering of agents, describing it as approximately equivariant. To ensure a fair comparison, the real dataset is ordered based on the initial positional role of each player.

In Table 2, we provide a comprehensive overview of metrics across methods solving the MP task. Our FootBots demonstrate superior performance in all metrics. The Velocity model shows a significant performance gap, attributed to unconstrained long-term predictions exceeding pitch boundaries. RNN [20] and FootBots NS, lacking agent interaction, lead to performance decline, mostly in the MaxErr$_P$ and FDE$_P$ metrics. This emphasizes the significance of social interaction in long-term trajectory prediction. In general, social-aware baselines like siMLPe [5] and baller2vec++ [26] achieve superior metrics compared to non-social methods. Although siMLPe competes well, MaxErr$_P$ and FDE$_P$ metrics are outperformed by transformer-based approaches like baller2vec++ and FootBots. Moreover, these methods can handle variable sequence length. However, baller2vec++ encounters difficulties with ball prediction, leading to suboptimal results due to error accumulation. FootBots excels in ball and players predictions, leverages permutation equivariance, and is more than six times faster than baller2vec++ in inference (73 vs 484 milliseconds), thanks to the decoupled attention. Importantly, in all MP results, ADE$_{ball}$ consistently exceeds ADE$_P$ across methods, motivating CMP$_1$ task.

Figure 4 provides an illustration of a particular trajectory prediction example, offering a comparative analysis of baselines in the MP task. Linear predictions based on Velocity baseline exhibit trajectories that are deemed non-sensical in certain instances. Both FootBots NS and RNN [20] models tend to generate shorter predicted trajectories, underscoring the imperative of incorporating social interaction

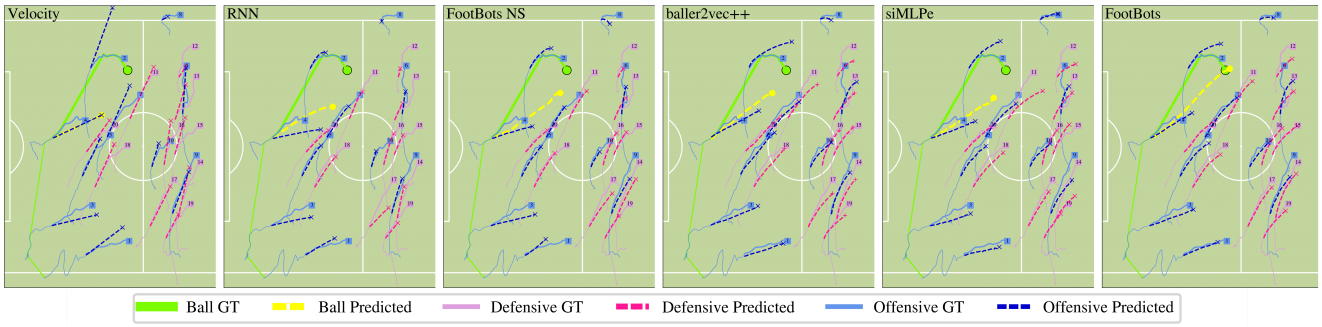


**Fig. 4**. **Qualitative evaluation and comparison on real data.** The figure displays the estimated trajectories for approaches Velocity, RNN [20], baller2vec++ [26] and siMLPe [5]; and our solutions FootBots NS and FootBots, by solving the MP task.

| Model | Order | Task | Predict ($P$) | $\text{ADE}_P \downarrow$ | $\text{ADE}_{\text{ball}} \downarrow$ | $\text{MaxErr}_P \downarrow$ | $\text{FDE}_P \downarrow$ | $\text{MR}_P$ (%) $\downarrow$ |
|---|---|---|---|---|---|---|---|---|
| Velocity | None | MP | Players+Ball | 3.27 | 9.39 | 7.34 | 7.27 | 67.50 |
| RNN [20] | None | MP | Players+Ball | 2.67 | 6.91 | 5.56 | 5.43 | 65.54 |
| FootBots NS (Ours) | None | MP | Players+Ball | 2.39 | 6.37 | 5.16 | 5.04 | 60.99 |
| baller2vec++ [26] | Approx-Equivariant | MP | Players+Ball | 2.21 | 6.43 | 4.64 | 4.49 | 60.79 |
| siMLPe [5] | Role-based | MP | Players+Ball | 2.18 | 6.15 | 4.73 | 4.55 | 59.71 |
| FootBots (Ours) | Equivariant | MP | Players+Ball | **2.04** | **5.79** | **4.43** | **4.28** | **57.37** |
| | | $\text{CMP}_1$ | Players | 1.64 | - | 3.42 | 3.20 | 52.59 |
| | | $\text{CMP}_2$ | Defensive | 1.38 | - | 2.80 | 2.59 | 47.83 |
| | | $\text{CMP}_3$ | Offensive | 1.44 | - | 2.98 | 2.78 | 48.26 |
| | | $\text{CMP}_4$ | Ball | 2.72 | 2.72 | 5.93 | 4.27 | 64.11 |

**Table 2**. **Quantitative evaluation and comparison for MP and CMP tasks on real data.** The table provides a comprehensive comparison of our solution with various other approaches in MP task. All metrics, except $\text{MR}_P$, are in meters.

for a comprehensive understanding of each player's future motions. Despite suboptimal ball prediction, siMLPe [5] and baller2vec++ [26] excels in accurate player predictions, showcasing its robustness in capturing player interaction dynamics. Additionally, FootBots outperforms previous baselines both quantitatively and qualitatively, with enhanced ball prediction.

To ensure the generalization of our findings, we conduct a parallel analysis using the real dataset to address all the considered CMP tasks. Similar to the synthetic dataset, we initiate the evaluation by assessing the model's performance in solving $\text{CMP}_1$. Our investigation also covers $\text{CMP}_2$, focusing on Defensive Players' positions, and $\text{CMP}_3$, targeting Offensive Players' ones. Furthermore, we explore $\text{CMP}_4$, which involves ball position prediction.

Quantitative results for FootBots across all CMP tasks are presented in Table 2. The solution for the $\text{CMP}_1$ task demonstrates marked improvement compared to the MP task, highlighting a strong correlation between player positions and the ball. Noteworthy enhancements in prediction accuracy emerge when conditioning on the opposing team and ball locations in $\text{CMP}_2$ and $\text{CMP}_3$. It is worth noting that predicting offensive team behaviors is more challenging than predicting defensive ones, due to their inherent stochasticity. Moreover, FootBots adeptly utilize player interactions to provide accurate ball predictions, reducing the $\text{ADE}_{\text{ball}}$ metric from 5.79 to 2.72 meters compared to the MP task.

Figure 1 illustrates a sample of the real dataset featuring solutions for the MP task and all CMP tasks. This specific instance depicts an scenario characterized by an extended ball trajectory involving a pass to predict and swift player motions. Notably, within the MP task, the model encounters challenges in accurately predicting both the ball and player positions, attributed to the inherent speed of the sequence. However, with the integration of conditioning information, discernible enhancements in predictions become evident.

## 5. CONCLUSIONS

In this work, we introduced FootBots, a tailored trajectory prediction model for soccer contexts, and extensively evaluated its performance across diverse scenarios. The comparative analysis demonstrated FootBots' superior performance over baseline models, showcasing its advantageous equivariance properties. Through synthetic dataset evaluation, FootBots excelled in predicting player positions, particularly when incorporating social attention and ball conditioning ($\text{CMP}_1$), highlighting the importance of social interactions and ball incorporation. Extension to real data showcased FootBots' grasp of defensive strategies ($\text{CMP}_2$), improved offensive player predictions ($\text{CMP}_3$), and effective player interaction utilization for ball position enhancement ($\text{CMP}_4$). Remarkably, $\text{CMP}_4$ exhibited significant error reduction compared to the MP task, affirming the effectiveness of using players as conditioning agents to enhance ball prediction accuracy.

## 6. REFERENCES


[1] O. B. Sezer, M. U. Gudelek, and A. M. Ozbayoglu, "Financial time series forecasting with deep learning: A systematic literature review: 2005–2019," *Applied Soft Computing*, vol. 90, pp. 106181, 2020. 1

[2] K. Fragkiadaki, S. Levine, P. Felsen, and J. Malik, "Recurrent network models for human dynamics," in *ICCV*, 2015. 1

[3] J. Martinez, M. J Black, and J. Romero, "On human motion prediction using recurrent neural networks," in *CVPR*, 2017. 1

[4] E. Aksan, M. Kaufmann, P. Cao, and O. Hilliges, "A spatio-temporal transformer for 3D human motion prediction," in *3DV*, 2021, pp. 565–574. 1

[5] W. Guo, Y. Du, X. Shen, V. Lepetit, X. Alameda, and F. Moreno-Noguer, "Back to mlp: A simple baseline for human motion prediction," in *WACV*, 2023. 1, 5, 6

[6] A. Alahi, K. Goel, V. Ramanathan, A. Robicquet, L. Fei-Fei, and S. Savarese, "Social LSTM: Human trajectory prediction in crowded spaces," in *CVPR*, 2016. 1

[7] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *CVPR*, 2018. 1

[8] V. Kosaraju, A. Sadeghian, R. Martín-Martín, I. Reid, H. Rezatofighi, and S. Savarese, "Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks," *NeurIPS*, 2019. 1

[9] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," in *ECCV*, 2020. 1

[10] J. Ngiam, V. Vasudevan, B. Caine, Z. Zhang, H. L. Chiang, J. Ling, R. Roelofs, A. Bewley, C. Liu, A. Venugopal, et al., "Scene transformer: A unified architecture for predicting future trajectories of multiple agents," in *ICLR*, 2021. 1, 3

[11] R. Girgis, F. Golemo, F. Codevilla, M. Weiss, J. A. D'Souza, S. E. Kahou, F. Heide, and C. Pal, "Latent variable sequential set transformers for joint multi-agent motion prediction," *ICLR*, 2022. 1, 3

[12] H. M Le, Y. Yue, P. Carr, and P. Lucey, "Coordinated multi-agent imitation learning," in *ICML*, 2017. 1, 2

[13] R. A Yeh, A. G Schwing, J. Huang, and K. Murphy, "Diverse generation for multi-agent sports games," in *CVPR*, 2019. 1, 2

[14] D. Ding and H H. Huang, "A graph attention based approach for trajectory prediction in multi-agent sports games," *arXiv:2012.10531*, 2020. 1, 2

[15] S. Hauri, N. Djuric, V. Radosavljevic, and S. Vucetic, "Multi-modal trajectory prediction of nba players," in *WACV*, 2021. 1, 2

[16] S. Omidshafiei, D. Hennes, M. Garnelo, Z. Wang, A. Recasens, et al., "Multiagent off-screen behavior prediction in football," *Scientific reports*, vol. 12, no. 1, pp. 8638, 2022. 1, 2

[17] M. A Alcorn and A. Nguyen, "baller2vec: A multi-entity transformer for multi-agent spatiotemporal modeling," *arXiv:2102.03291*, 2021. 1, 2

[18] M. Teranishi, K. Tsutsui, K. Takeda, and K. Fujii, "Evaluation of creating scoring opportunities for teammates in soccer via trajectory prediction," in *MLSA*, 2022. 1

[19] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, "Attention is all you need," *NeurIPS*, 2017. 1, 3

[20] S. Becker, R. Hug, W. Hubner, and M. Arens, "Red: A simple but effective baseline predictor for the trajnet benchmark," in *ECCVW*, 2018. 1, 5, 6

[21] Y. Yuan, X. Weng, Y. Ou, and K. M Kitani, "Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting," in *ICCV*, 2021. 1

[22] P. Felsen, P. Lucey, and S. Ganguly, "Where will they go? predicting fine-grained adversarial multi-agent motion using conditional variational autoencoders," in *ECCV*, 2018. 2

[23] E. Zhan, S. Zheng, Y. Yue, L. Sha, and P. Lucey, "Generating multi-agent trajectories using programmatic weak supervision," *arXiv:1803.07612*, 2018. 2

[24] S. Zheng, Y. Yue, and J. Hobbs, "Generating long-term trajectories using deep hierarchical networks," *NeurIPS*, 2016. 2

[25] C. Sun, P. Karlsson, J. Wu, J. B Tenenbaum, and K. Murphy, "Stochastic prediction of multi-agent interactions from partial observations," *arXiv:1902.09641*, 2019. 2

[26] M. A Alcorn and A. Nguyen, "baller2vec++: A look-ahead multi-entity transformer for modeling coordinated agents," *arXiv:2104.11980*, 2021. 2, 5, 6

[27] J. Lee, Y. Lee, J. Kim, A. Kosiorek, S. Choi, and Y. W. Teh, "Set transformer: A framework for attention-based permutation-invariant neural networks," in *ICML*, 2019. 3