Event2Tracking: Reconstructing Multi-Agent Soccer Trajectories Using Long-Term Multimodal Context

Harry Hughes[1,2], Michael Horton[1], Xinyu Wei[1], Harshala Gammulle[2], Clinton Fookes[2], Sridha Sridharan[2], Patrick Lucey[1]

[1]Stats Perform

[2]Queensland University of Technology

{harry.hughes, michael.horton, xinyu.wei, patrick.lucey}@statsperform.com,

{pranali.gammule, c.fookes, s.sridharan}@qut.edu.au

Abstract

Soccer is a rich testbed for studying multi-agent adversarial systems. In this work we focus on the task of reconstructing the noisy trajectories of soccer agents (players and the ball). Previous works that model the behaviours of agents in soccer are limited in two respects: (i) they only focus on short-term context windows ($\leq$10 seconds) which are not suitable for reconstructing trajectories impacted by long-term noise, and (ii) they exclusively rely on trajectory context, and do not leverage soccer's auxiliary data streams that can provide additional context. Our Event2Tracking model addresses these limitations. First, our architecture models soccer's *long-term* structure by processing long-term trajectories (60 seconds in duration). Secondly, our architecture is *multimodal*. Specifically, it fuses soccer tracking data with event data (which specifies the high-level semantic events that transpire in a game), providing rich context that cannot strictly be inferred from the raw trajectories. We evaluate our method empirically using a reconstruction loss metric. Compared to state-of-the-art approaches, our method substantially improves the accuracy of the ball's and players' reconstructed trajectories.

Introduction

The behaviours of agents (players and the ball) in soccer form a rich and important testbed for the study of multi-agent adversarial systems (Yeh et al. 2019; Tuyls et al. 2021; Omidshafiei et al. 2022; Wang et al. 2024). In this paper, we model the fine-grained spatiotemporal behaviours of agents in professional soccer games. The availability of data which encodes agents' fine-grained spatiotemporal behaviours is a fundamental prerequisite for modelling soccer games. One such data stream is multi-agent tracking data, which specifies each agent's 2D centre of mass at a high framerate ($\sim$25Hz). Multi-agent tracking data is typically generated using computer vision systems[1] that are installed *in-venue*. However, the prohibitive cost of these systems limit their broad adoption. A scalable alternative to in-venue systems is *broadcast tracking*, where agents are tracked remotely using computer vision from publicly accessible broadcast footage.

[1]Computer vision is preferred to wearable technology (i.e., RFID or GPS) as it is unobtrusive and less susceptible to hardware failure.

Reconstructing soccer's broadcast tracking poses many challenges from a modelling perspective. First, players in broadcast footage frequently exit and enter the moving camera's field-of-view, resulting in heavy occlusions. Although occluded players are outside the camera's receptive field, they are still active in the game i.e., they adhere to structured individual roles, while still responding to the behaviours of their teammates and opponents. The need to model long-term off-screen behaviours differentiates soccer from other frequently studied multi-agent tracking scenes. For example, in pedestrian environments, agents that are outside the camera's field-of-view are not typically modelled and are assumed to be irrelevant to the scene. Additionally, broadcast cameras in other invasion games like American football and basketball typically have much wider fields-of-view relative to the size of the area-of-interest. This results in much shorter-term occlusions in these games.

Another challenge lies in reconstructing the trajectory of the ball. The purpose of soccer is to score goals, which occurs when the ball crosses either team's goal-line. This makes the ball the focal point of soccer. Even though building a computer vision-based ball detector is quite a trivial task, its small size, fast movement, heavy occlusion from players, and visual similarity to other entities on the pitch (e.g., pitch markings, players' boots) make the ball extremely difficult to accurately and continuously track from broadcast footage. Previous impressive works that model soccer scenes (Hoshen 2017; Le et al. 2017; Yeh et al. 2019; Omidshafiei et al. 2022) are limited in two respects. First, they only focus on short-term trajectories (typically $\leq$10 seconds in duration), and therefore do not model the game's longer-term dynamics. Secondly, they model soccer scenes unimodally (only using trajectory context). This is especially limiting when reconstructing the motion of the ball, as its location must be inferred entirely from the motion of players. This task becomes profoundly difficult in periods of heavy occlusion.

We tackle these two limitations in our work, presenting the Event2Tracking architecture, which is a long-term multimodal trajectory reconstruction model identify spatiotemporal axial attention (Monti et al. 2022; Nayakanti et al. 2023) as an effective approach to model longer trajectories than previously studied (60 seconds in duration rather than $\leq$10 seconds). We also suggest an elegant method for jointly

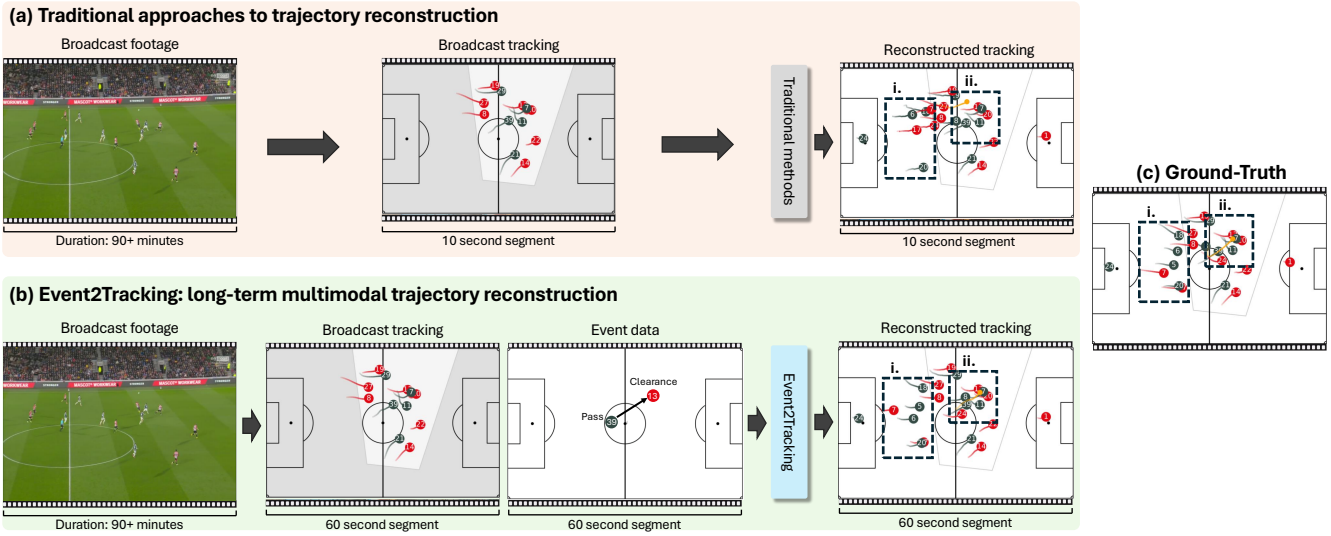

Figure 1: **Compares traditional approaches to trajectory reconstruction to our approach**. We focus on reconstructing multi-agent tracking data extracted from broadcast footage i.e., broadcast tracking. Traditional approaches (a) reconstruct trajectories using short-term ($\leq$10 seconds) unimodal trajectory context. In our Event2Tracking model (b), we use long-term trajectories (60 seconds) as well as soccer event data, which specifies the semantic sequence of high-level actions that transpire across the game). Compared to traditional approaches, our method results in reconstructed trajectories that more closely match the ground-truth (c). This is most evident in terms of the trajectories of occluded players (as is shown in i.) and the motion of the ball (as is shown in ii.)

modelling long-term trajectories and *event data*. Event data is a sparse spatiotemporal data stream which specifies the location, timestamp, and identity of each on- and off-ball event in the game. This information stream is labelled at-scale and reliably by human annotators[2]. As demonstrated in the Experiments Section, this long-term multimodal context substantially increases the accuracy of the ball and players' reconstructed motion. A comparison between our approach and traditional trajectory modelling approaches in sport is visualised in Figure 1. In summary, our contributions are as follows:

- We demonstrate that spatiotemporal axial attention (Monti et al. 2022; Nayakanti et al. 2023) is an effective approach for modelling longer trajectories than previously studied (60 seconds in duration rather than $\leq$10 seconds).
- We present an elegant method for jointly modelling long-term trajectories with soccer's event data.
- We compare against state-of-the-art baselines on the task of reconstructing soccer broadcast tracking, showing that our approach is able to more accurately reconstruct the trajectories of players and the ball.

[2]Humans remain more accurate than automated event detectors (Deliege et al. 2021; Vidal-Codina et al. 2022), which is crucial for media and betting purposes.

Related Work

Modelling Multi-Agent Trajectories

We start by detailing existing methods for modelling multi-agent trajectories, focusing on two environments which consist of multiple humans interacting in a continuous spatiotemporal environment: pedestrian scenes and sporting scenes.

In Pedestrian Scenes Seminal early works in pedestrian trajectory prediction use heuristic and energy-based methods to model agents' spatiotemporal relationships (Helbing and Molnar 1995; Wang, Hertzmann, and Fleet 2005; Van den Berg, Lin, and Manocha 2008; Pellegrini et al. 2009). Deep learning methods have been shown to be well-suited to extracting the non-linear multi-agent dynamics from tracking data. Recurrent neural networks (RNNs) are frequently used to model each agent's temporal history. This temporal context is typically distributed spatially via pooling (Alahi et al. 2016; Gupta et al. 2018; Wang et al. 2023) or with graph neural networks (GNNs) (Zhang et al. 2019; Salzmann et al. 2020). With the success of Transformers (Vaswani et al. 2017) in sequential learning tasks (Devlin et al. 2018; Brown et al. 2020; Jumper et al. 2021), attention-based architectures are now used to jointly encode both the spatial and temporal dimensions of multi-agent trajectories (Yuan et al. 2021; Giuliari et al. 2021; Xia et al. 2021; Zhou et al. 2022). However, Transformers have quadratic complexity with respect to sequence length, which is especially limiting when applied to high-dimensional multi-agent trajectory sets. As a result, recent works aim to in-

crease the efficiency of Transformers when applied to tracking data. One notable approach is spatiotemporal axial attention (Monti et al. 2022; Nayakanti et al. 2023), which applies self-attention separately across the temporal and spatial axes of multi-agent trajectory sets.

These approaches typically focus on short-term trajectories ($\leq$10 seconds in duration). This is because (i) these trajectories are gathered using cameras with relatively narrow fields-of-view, and (ii) off-screen behaviours are assumed to not be relevant to scenes. Despite this, we observe that spatiotemporal axial attention has suitable properties for modelling longer trajectories than previously studied.

In Sporting Scenes Several notable works have modelled multi-agent trajectories in sporting scenes, where the most widely researched setting is that of trajectory forecasting over short-term horizons ($\leq$10 seconds) (Zheng, Yue, and Lucey 2016; Le et al. 2017; Hoshen 2017; Felsen, Lucey, and Ganguly 2018; Zhan et al. 2018; Yeh et al. 2019). Multiple works also study multi-agent trajectory imputation in sporting scenes. For example, (Liu et al. 2019) use bidirectional context to impute missing basketball trajectories. However, this approach models each agent independently, and does not model the spatial correlations that exist in multi-agent scenes. While (Xu et al. 2023) does model these spatial correlations, they only leverage past temporal context. Most closely related to our research setting is the *Graph Imputer* (Omidshafiei et al. 2022), which also focuses on reconstructing soccer broadcast tracking data using bidirectional temporal context. This approach models bidirectional context by making two independent predictions, one operating forwards in time (only using past context) and one operating backwards in time (only using future context). These directional predictions follow (Yeh et al. 2019) and are fused via averaging. Separately modelling future and past context is more limited for longer trajectories, where forwards and backwards predictions tend to be less closely correlated. However, in the original work they only focus on short-term trajectories (9.6 seconds) where the first and final seconds are visible. As a result, this limitation is not encountered.

In contrast to (Omidshafiei et al. 2022), we investigate a more realistic setting for the reconstruction of broadcast tracking. Specifically, as we use real broadcast tracking data, we make no assumptions about the visibility of agents (e.g., at the starts or ends of trajectories). This considerably increases both the duration of agent occlusions, and as a result, the difficulty of the trajectory reconstruction task.

Multimodal Context in Trajectory Modelling

In many environments, the behaviours of agents strongly depends on scene-level context. One extensive line of work is in extracting static map elements from top-down images of scenes using convolutional feature extractors (Kosaraju et al. 2019; Sadeghian et al. 2019; Casas et al. 2020; Fang et al. 2020; Phan-Minh et al. 2020). These approaches are limited by (i) the high dimensionality of convolutional feature maps which makes modelling longer sequences difficult, and (ii) the need for complex handcrafted fusion of image features with multi-agent trajectories. Recent works have shown the Transformer's broad utility in fusing diverse data modalities such as text, video, and audio (Jaegle et al. 2021; Radford et al. 2021; Alayrac et al. 2022; Girdhar et al. 2023). Following this trend, the Wayformer (Nayakanti et al. 2023) uses attention-based architectures to encode and fuse multi-agent trajectories with other spatiotemporal modalities relevant in an autonomous driving setting. In another research stream, (Everett et al. 2023) exclusively uses soccer's event stream to infer the locations of agents at each event (using no trajectory context).

Inspired by (Nayakanti et al. 2023; Everett et al. 2023), we fuse soccer event data and multi-agent trajectories using a Transformer-based representation.

Proposed Method

Problem Formulation

For our trajectory reconstruction setting, we have access to two tracking streams: broadcast tracking (which contains occlusions and noise) and in-venue tracking (which is complete and accurate). Broadcast tracking for E agents over T timesteps can be represented as a spatiotemporal grid $\boldsymbol{m}_{\text{broadcast}} \in \mathbb{R}^{T \times E \times d_{\text{broadcast}}}$. Each observation in broadcast tracking contains $d_{\text{broadcast}}$ features, which include the agent's 2D coordinates, and one-hot encodings of the agent's role, their team affiliation, and their team's current formation. When trajectory observations are occluded, the agent's (x, y) location is set to a constant value outside the pitch's coordinates. The in-venue stream $\boldsymbol{m}_{\text{in-venue}} \in \mathbb{R}^{T \times E \times 2}$ consists of each agent's (x, y) location at each timestep in the trajectory. Event data is a 1D temporal stream $\boldsymbol{m}_{\text{event}} \in \mathbb{R}^{L \times d_{\text{event}}}$ where L is the number of events in trajectory window and d_{event} is the dimensionality of each event observation. Each event token includes the 2D coordinate of the event, and one-hot encodings of the event type (e.g., pass, shot, control), and the focused agent's team affiliation, role, and their team's current formation. The training objective is to learn a function F parameterised by θ^* where

$$\theta^* := \min_{\theta} \mathcal{L}_2(F_\theta(\boldsymbol{m}_{\text{broadcast}}, \boldsymbol{m}_{\text{event}}), \boldsymbol{m}_{\text{in-venue}}).$$

Encoding Long Multi-Agent Trajectories

Agents dynamically enter and exit the broadcast camera's field-of-view. However, despite being off-screen, occluded agents are still relevant to the scene. They have structured long-term roles, and constantly evolving behaviours based on the actions of their teammates and opposition. As we show experimentally, longer contexts of up to 60 seconds improve the capacity to reconstruct these impeded trajectories.

One approach for efficiently modelling multi-agent trajectories with self-attention is spatiotemporal axial attention (Monti et al. 2022; Nayakanti et al. 2023). Spatiotemporal axial attention is a module where self-attention is applied across the temporal and spatial axes of multi-agent trajectory sets separately. With this scheme, individual agent motion can be learned through temporal attention, while collective group dynamics can be learned through spatial attention. This is illustrated in Figure 2. A key benefit of

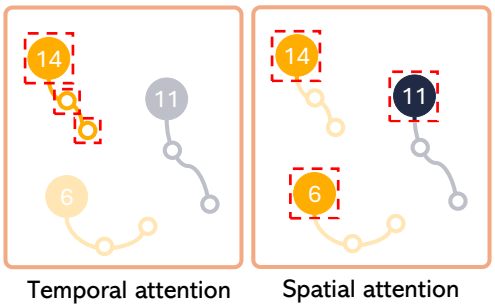

Figure 2: Illustration of spatiotemporal axial attention. In both attention modules, the red squares denote the tokens used for a single attention computation. In temporal attention, self-attention is applied independently within each agent's trajectory, modelling each agent's temporal context. Spatial attention applies self-attention within each individual timestep, modelling the inter-agent spatial dependencies that exist in the environment.

spatiotemporal axial attention is its computation efficiency. Self-attention has quadratic performance with respect to sequence length. Therefore, jointly attending across spatial and temporal axes of trajectories has $\mathcal{O}(T^2 \cdot E^2)$ complexity. As is noted in (Nayakanti et al. 2023), separate axial attention is of complexity $\mathcal{O}(T^2) + \mathcal{O}(E^2) = \mathcal{O}(T^2)$ where sequence length T dominates the number of agents E. Despite the efficiency of spatiotemporal axial attention, it has previously only been applied to short-term trajectories ($\leq$10 seconds in duration). In this work, we make the further observation that the efficiency of spatiotemporal axial attention makes it suitable for modelling considerably longer-term behaviours.

Spatiotemporal axial attention also enables processing of multi-agent trajectories without imposing an artificial ordering on agents. While spatiotemporal data has a clear temporal total ordering (i.e., chronological), no such natural ordering exists over agents spatially. In soccer, because there are two teams with 10 outfield players, there are $(10!)^2$ possible permutations of agent indices. Consequently multi-agent trajectory sets must be modelled in a way that is permutation equivariant to avoid a combinatorial increase in complexity. Previous approaches handle this by imposing an artificial ordering on players based on their locations (Lucey et al. 2013; Sha et al. 2017). Instead, spatiotemporal axial attention processes multi-agent trajectories in a natively permutation equivariant manner. That is, when modelling trajectories $\boldsymbol{m}_{\text{broadcast}}$ with a function which uses spatiotemporal axial attention f, the following equality holds:

$$f(\boldsymbol{m}_{\text{broadcast}})^p = f(\boldsymbol{m}^p_{\text{broadcast}}),\ \forall\ p \in [1, (10!)^2], \qquad (1)$$

where p represents a permutation of the agent indices in the output of spatiotemporal axial attention function $f(\boldsymbol{m}_{\text{broadcast}})$ and the broadcast tracking input $\boldsymbol{m}_{\text{broadcast}}$.

Event2Tracking Model

This section details the Event2Tracking model, which uses spatiotemporal axial attention as a core operation. We first outline our method for temporal localisation, before outlining our event encoder and tracking decoder architectures (shown in Figure 3).

Temporal Localisation One task common to both the encoding event and tracking data is that of *temporal localisation*. That is, specifying the exact timing of each event and tracking observation. The central challenge here is that both input data sources have non-uniform time intervals (broadcast tracking data is generated at a variable frame-rate, and events occur sparsely). To address this, for each token, we calculate the time elapsed (in milliseconds) from the start of the current trajectory window. We use this integer value as the index used for sinusoidal positional encoding (Vaswani et al. 2017), allowing for flexible encoding of time in both of our multimodal inputs.

Event Encoder The second component of the Event2Tracking model is the event encoder. This module encodes each event in $\boldsymbol{m}_{\text{event}}$. Events are first tokenised through linear projection, before adding sinusoidal positional embeddings to specify each event's temporal occurrence (as detailed above). These tokens are then processed by a vanilla Transformer encoder with N layers, producing event embeddings $\boldsymbol{z}_{\text{event}} \in \mathbb{R}^{L \times d_h}$ of latent dimensionality d_h.

Tracking Decoder Our tracking decoder is heavily inspired by spatiotemporal axial attention (Monti et al. 2022; Nayakanti et al. 2023). However, our architecture enables the joint modelling of event encodings $\boldsymbol{z}_{\text{event}}$ with broadcast tracking data $\boldsymbol{m}_{\text{broadcast}}$. Each tracking observation is first tokenised through a linear projection. Following this, sinusoidal positional embeddings are added to specify the temporal ordering of trajectory tokens (as outlined in the section above). Tokens are then encoded by an attention-based module that is stacked N times. Within this module, tokens are first processed by spatiotemporal axial attention (temporal attention followed by spatial attention). This encodes the spatiotemporal dependencies of multi-agent trajectories in an efficient, permutation equivariant manner. Next, each agent's trajectory tokens are cross-attended with the event embeddings $\boldsymbol{z}_{\text{event}}$ independently. This temporal cross-attention operation fuses broadcast tracking tokens with event context. Next are the normalisation and feedforward layers standard to Transformers (Vaswani et al. 2017). The tracking decoder model returns $\boldsymbol{z}_{\text{broadcast}} \in \mathbb{R}^{T \times E \times d_h}$, which represents joint encodings of each agent's event and broadcast tracking streams. Finally, a linear projection is used to map each token to an (x, y) prediction.

Experiments

Experimental Setup

Dataset A large dataset was used in the experiments, with 700 professional soccer games for training and 52 games for evaluation. Each game had a paired dataset of event data $\boldsymbol{m}_{\text{event}}$, broadcast tracking $\boldsymbol{m}_{\text{broadcast}}$, and in-venue tracking $\boldsymbol{m}_{\text{in-venue}}$. The ball's trajectory is considered to be fully occluded in the broadcast tracking dataset, representing the challenges in tracking the ball continuously and accurately

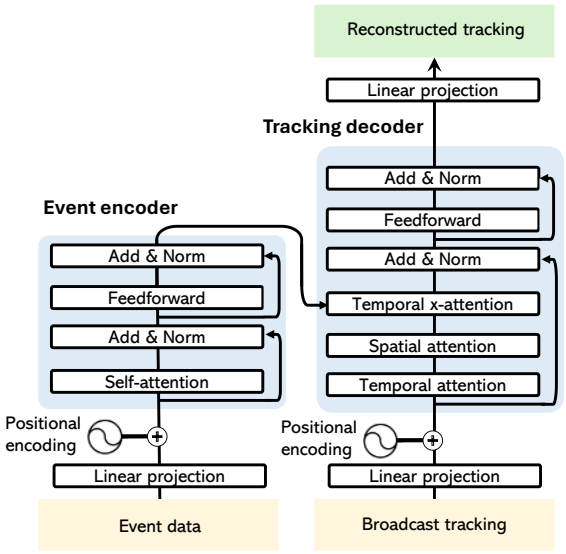

Figure 3: Illustration of our Event2Tracking architecture, which jointly models event data and broadcast tracking data to reconstruct multi-agent trajectory sets.

throughout the entire game using a heavily occluded monocular camera. An inventory and justification of this dataset is provided in Appendix 1.

Metric A reconstruction loss metric is used for quantitative evaluation. Specifically, average displacement error (ADE) is used, which computes the average euclidean distance (m) between reconstructed and real locations within a certain trajectory segment. We report mean ADE (mADE), which takes the mean ADE calculated over 1-minute trajectory segments both for players and the ball.

Baselines We focused on evaluating our method for modelling soccer scenes with bidirectional temporal context. As a result, although multi-agent sporting trajectories are inherently stochastic (Omidshafiei et al. 2022; Yeh et al. 2019; Sun et al. 2019), we only evaluated our method against deterministic baselines.

- **Linear interpolator**: interpolates behaviours between available observations in broadcast tracking. Where players are not visible over the entire trajectory window, their locations are set to the centroid of their team's locations.
- **Independent Transformer** (Vaswani et al. 2017): reconstructs each agent trajectory independently using a Transformer.
- **Graph Imputer** (Omidshafiei et al. 2022): reconstructs trajectories by averaging predictions made forwards and backwards in time. Each directional prediction uses a RNN to model each agent's temporal context, before distributing this context via a GNN. We ablated the original method's stochasticity.
- **Spatiotemporal Transformer** (STT) (Monti et al. 2022): uses a Transformer with spatiotemporal axial attention. While the original method only uses past context, we enable bidirectional context by removing the autoregressive attention mask.

Implementation Details All models were trained separately using 10, 20, 30, 45, and 60 second context windows to quantify how each approach generalised to longer trajectories. Trajectories of greater length were not considered due to computational constraints. The broadcast and in-venue tracking streams were downsampled to 5Hz. Each attention module uses a hidden dimensionality of 128, and a feedforward dimensionality of 512 and 4 attention heads. For the Event2Tracking module, the event encoder and tracking decoder each have $N = 4$ layers. During training, the loss incurred in prediction the ball location was weighted by a factor of 11, reflecting the ball's relative importance. All models were trained for 16 hours on a a cluster of 4 A10 GPUs with a learning rate of $1e-4$ using the Adam optimiser (Kingma and Ba 2014) (with default exponential decay parameters).

Quantitative Results

We start by quantitatively comparing our approach to each baseline when trained on different segment lengths (10s, 20s, 30, 45s, 60s). The mADE reconstruction loss metrics are shown for the players and ball in Table 1. Notably, our proposed Event2Tracking architecture outperforms all baselines over every segment length investigated.

The first trend we observe is that the Event2Tracking model has the strongest performance in terms of reconstructing the ball's motion. Specifically, our method outperforms the next best model (STT) by between 32% and 36% in terms of mADE (ball) across every context length. These architectures use identical methods to encode the broadcast tracking data (spatiotemporal axial attention). However, recall that the ball's trajectory is fully occluded in broadcast tracking. As a result, unimodal methods (such as the STT) must infer the ball's trajectory only using the motion of visible players. In contrast, our method uses event data, which contains the time, location, and player identity of every on-ball event in the game. Our results indicate that this auxiliary information source is beneficial when predicting the ball's location.

The Event2Tracking model also has the best performance in terms of reconstructing player locations. Our method shows between 3% and 11% lower mADE (players) values across each context window length than the next best model (STT). This is logical, as event data also provides spatiotemporal context pertaining to the locations of players i.e., it provides the location of players when they complete an event. While these improvements are lower in magnitude than the improvements in terms of reconstructing the ball, they further reinforce the utility that event data provides when reconstructing heavily impeded trajectories.

Next, of the deep learning methods, our approach shows the strongest performance improvements when applied to longer context windows. The Event2Tracking's mADE (players) monotonically improves when applied to longer trajectories (Figure 4). Specifically, its performance for this metric improves 22% between 10 and 60 second context

Context	mADE Player/Ball (m)				
	Linear interpolator	Independent Transformer	Graph Imputer	Spatiotemporal Transformer	Event2Tracking (ours)
10s	8.98/-	5.80/17.81	4.66/7.69	4.25/6.23	**4.13/4.24**
20s	7.88/-	5.30/17.27	4.45/7.56	3.81/5.71	**3.44/3.76**
30s	7.35/-	5.22/16.96	4.41/7.56	3.64/5.33	**3.33/3.52**
45s	6.95/-	4.77/16.63	4.43/7.73	3.62/5.48	**3.27/3.53**
60s	6.72/-	4.78/16.69	4.60/7.99	3.62/5.46	**3.22/3.51**

Table 1: Compares our method to the baselines when trained on segments of different lengths (10s, 20s, 30s, 45s, 60s).

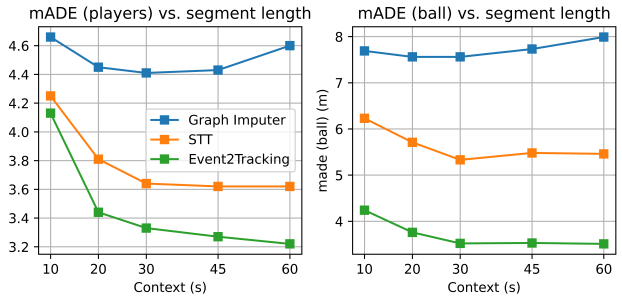

Figure 4: **Reconstruction loss over longer segments**. Our Event2Tracking outperforms each baseline over each context window, and tends to improve in performance with more temporal context.

windows. A similar trend can be observed in the STT's performance (which has the next-best performance), which improves 15% between 10 and 60 second context windows. This is a meaningful result, strongly indicating that spatiotemporal axial attention is an effective method for modelling long-term trajectories. Additionally, it highlights the importance of modelling long-term context when reconstructing heavily impeded soccer trajectories. In contrast, the Graph Imputer's performance only improves 5% from 10s to 30s, before decreasing when applied to longer segments. Its performance is also the weakest of these three models over every segment length. This result highlights the limitations of the Graph Imputer for modelling long-term bidirectional context.

To make these results more concrete, our model's performance over a single representative game can be inspected. In Figure 5, we compare the Event2Tracking (60s) against the STT (10s) and (60s) baselines. In terms of mADE (players), the Event2Tracking model has strictly lower values than both baselines over the entire game. Our method also has the strongest performance over in terms of mADE (ball). While the STT (60s) outperforms our method in eight of the forty-eight 1-minute intervals for this metric, we suggest that this is expected variance due to the ball's fast and volatile movement. Additionally, the Event2Tracking's mADE values for both metrics are much more robust for this game, showing fewer high outlier values. This game further emphasises the efficacy of our Event2Tracking approach.

Finally, we note that the weakest performance is exhibited by the Linear Interpolator and Independent Transformer. Recall that the ball's trajectory is fully occluded in broadcast tracking. As a result, the Linear Interpolator is unable to reconstruct its trajectory. This highlights a limitation of interpolation-based approaches. Another limitation of these models is that they process each agent's trajectory independently. The impact of this is especially clear in terms of the Independent Transformer's high ball mADE value. As the ball has no detections, its motion can only be inferred from other agents' motion, or additional streams of information (i.e., event data). As the Independent Transformer does not model either, it is unable to accurately reconstruct the ball's trajectory. The inability to model inter-agent dependencies also results in these models having the two highest mADE (player) metrics for every context window. These results highlight the importance of modelling inter-agent dependencies in reconstructing soccer tracking data.

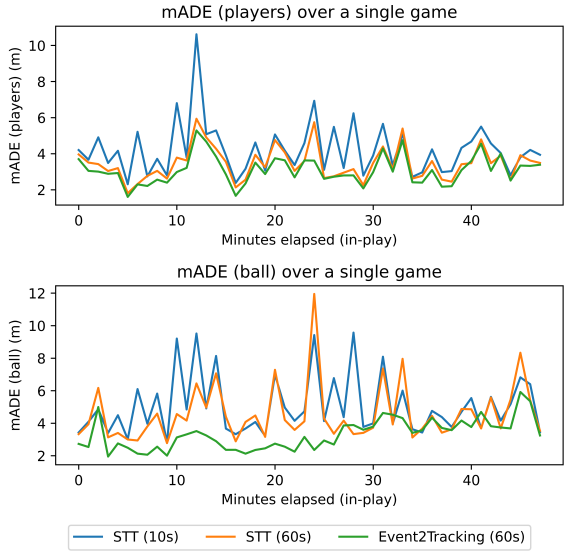

Figure 5: **Performance over a full game for when the ball is in-play**[3]. We show mADE for players (top) and the ball (bottom). Our Event2Tracking model (60s) outperforms both approaches in terms of both metrics for the vast majority of the game.

[3]While soccer games last for 90+ minutes, the ball is only in-play 50-60% of this time (Hopkins 2023).

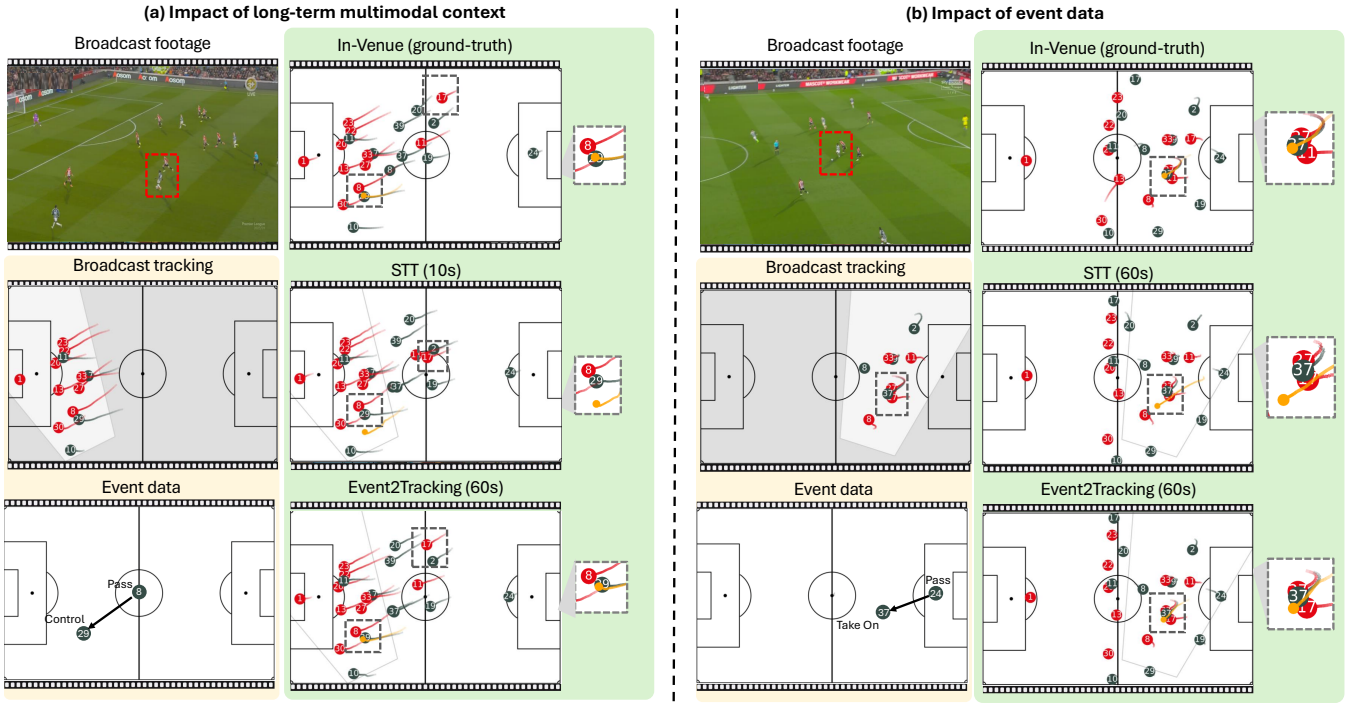

Figure 6: **Qualitative results**, where the broadcast footage, broadcast tracking, event data, in-venue tracking, baseline reconstruction, and Event2Tracking reconstructions are shown: (a) shows a frame where green player #29 completes a control event, (b) shows a frame where green #37 player completes a take-down event (where they attempt to dribble past an opponent). Our method's reconstructions more closely resemble in-venue compared to baselines.

Qualitative Results

Figure 6a shows a scene where green #29 is in control of the ball. In broadcast tracking, nine of the twenty-two players are occluded. The Event2Tracking (60s) and STT (10s) reconstructions are different in two key respects. First, the Event2Tracking predicts the location of green #17 much more accurately than the STT. This agent is not visible in broadcast footage for the 10 seconds surrounding this individual frame. As a result, only models that use >10 are able to attend to this player's past and future trajectories. This highlights the importance of using long-term context when reconstructing long-term occlusions. The second major difference is in the ball's trajectory. The STT (10s) baseline predicts the ball to be multiple metres away from the green #29 (who completes the control event). This is both visibly different to in-venue and represents an unrealistic soccer behaviour; a control event cannot occur if a player in not in possession of the ball. This highlights the challenges of predicting the ball's location from player motion alone. In contrast, our Event2Tracking method uses the motion of surrounding players as well as event data, which the time, location, and player involved in the control event. Consequently, the Event2Tracking's predicted ball trajectory closely resembles the in-venue ball trajectory. This example emphasises the importance of leveraging event context, especially in reconstructing the ball's trajectory.

The importance of using event data is reinforced in Figure 6b, where green #37 completes a take-on event i.e., where a player attempts to dribble past an opponent. In this example, the STT (60s) predicts the ball to be multiple metres away from the player completing the event. As in the example above, this is unrealistic in soccer, as a take-on event requires the player to be in possession of the ball. In contrast, Event2Tracking approach predicts the ball's location to be in close proximity to green #37's location, closely resembling the in-venue tracking. This examples further highlight the utility that event data provides.

Conclusion and Future Work

We outline a method for reconstructing heavily impeded multi-agent soccer trajectories. We illustrate that long-term multimodal context improves reconstructions of noisy trajectories. This is shown experimentally, where we compare against multiple approaches from previous work. In terms of future work, an exciting direction is combining our Event2Tracking architecture with a generative model, which would enable the stochastic, diverse, and controllable generation of behaviours that are also consistent with soccer's multimodal long-term structure. Additionally, given that our architecture has been shown to effectively model soccer scenes, another interesting avenue is to utilise this model as a general-purpose architecture for detecting and predicting other team and player behaviours (e.g., likelihood of a team scoring a goal within a certain time-horizon).

References

Alahi, A.; Goel, K.; Ramanathan, V.; Robicquet, A.; Fei-Fei, L.; and Savarese, S. 2016. Social lstm: Human trajectory prediction in crowded spaces. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 961–971.

Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35: 23716–23736.

Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J. D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33: 1877–1901.

Casas, S.; Gulino, C.; Liao, R.; and Urtasun, R. 2020. Spagnn: Spatially-aware graph neural networks for relational behavior forecasting from sensor data. In *2020 IEEE International Conference on Robotics and Automation (ICRA)*, 9491–9497. IEEE.

Deliege, A.; Cioppa, A.; Giancola, S.; Seikavandi, M. J.; Dueholm, J. V.; Nasrollahi, K.; Ghanem, B.; Moeslund, T. B.; and Van Droogenbroeck, M. 2021. Soccernet-v2: A dataset and benchmarks for holistic understanding of broadcast soccer videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 4508–4519.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Everett, G.; Beal, R. J.; Matthews, T.; Early, J.; Norman, T. J.; and Ramchurn, S. D. 2023. Inferring Player Location in Sports Matches: Multi-Agent Spatial Imputation from Limited Observations. *arXiv preprint arXiv:2302.06569*.

Fang, L.; Jiang, Q.; Shi, J.; and Zhou, B. 2020. Tpnet: Trajectory proposal network for motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6797–6806.

Felsen, P.; Lucey, P.; and Ganguly, S. 2018. Where will they go? predicting fine-grained adversarial multi-agent motion using conditional variational autoencoders. In *Proceedings of the European conference on computer vision (ECCV)*, 732–747.

Girdhar, R.; El-Nouby, A.; Liu, Z.; Singh, M.; Alwala, K. V.; Joulin, A.; and Misra, I. 2023. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 15180–15190.

Giuliari, F.; Hasan, I.; Cristani, M.; and Galasso, F. 2021. Transformer networks for trajectory forecasting. In *2020 25th international conference on pattern recognition (ICPR)*, 10335–10342. IEEE.

Gupta, A.; Johnson, J.; Fei-Fei, L.; Savarese, S.; and Alahi, A. 2018. Social gan: Socially acceptable trajectories with generative adversarial networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2255–2264.

Helbing, D.; and Molnar, P. 1995. Social force model for pedestrian dynamics. *Physical review E*, 51(5): 4282.

Hopkins, O. 2023. The Definitive Guide to Premier League Time-Wasting. https://theanalyst.com/na/2023/05/guide-to-premier-league-time-wasting/. Accessed: 2024-08-14.

Hoshen, Y. 2017. Vain: Attentional multi-agent predictive modeling. *Advances in neural information processing systems*, 30.

Jaegle, A.; Gimeno, F.; Brock, A.; Vinyals, O.; Zisserman, A.; and Carreira, J. 2021. Perceiver: General perception with iterative attention. In *International conference on machine learning*, 4651–4664. PMLR.

Jumper, J.; Evans, R.; Pritzel, A.; Green, T.; Figurnov, M.; Ronneberger, O.; Tunyasuvunakool, K.; Bates, R.; Žídek, A.; Potapenko, A.; et al. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873): 583–589.

Kingma, D. P.; and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Kosaraju, V.; Sadeghian, A.; Martín-Martín, R.; Reid, I.; Rezatofighi, H.; and Savarese, S. 2019. Social-bigat: Multimodal trajectory forecasting using bicycle-gan and graph attention networks. *Advances in Neural Information Processing Systems*, 32.

Le, H. M.; Yue, Y.; Carr, P.; and Lucey, P. 2017. Coordinated multi-agent imitation learning. In *International Conference on Machine Learning*, 1995–2003. PMLR.

Liu, Y.; Yu, R.; Zheng, S.; Zhan, E.; and Yue, Y. 2019. Naomi: Non-autoregressive multiresolution sequence imputation. *Advances in neural information processing systems*, 32.

Lucey, P.; Bialkowski, A.; Carr, P.; Morgan, S.; Matthews, I.; and Sheikh, Y. 2013. Representing and discovering adversarial team behaviors using player roles. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2706–2713.

Monti, A.; Porrello, A.; Calderara, S.; Coscia, P.; Ballan, L.; and Cucchiara, R. 2022. How many observations are enough? knowledge distillation for trajectory forecasting. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 6553–6562.

Nayakanti, N.; Al-Rfou, R.; Zhou, A.; Goel, K.; Refaat, K. S.; and Sapp, B. 2023. Wayformer: Motion forecasting via simple & efficient attention networks. In *2023 IEEE International Conference on Robotics and Automation (ICRA)*, 2980–2987. IEEE.

Omidshafiei, S.; Hennes, D.; Garnelo, M.; Wang, Z.; Recasens, A.; Tarassov, E.; Yang, Y.; Elie, R.; Connor, J. T.; Muller, P.; et al. 2022. Multiagent off-screen behavior prediction in football. *Scientific reports*, 12(1): 8638.

Pellegrini, S.; Ess, A.; Schindler, K.; and Van Gool, L. 2009. You'll never walk alone: Modeling social behavior for multi-target tracking. In *2009 IEEE 12th international conference on computer vision*, 261–268. IEEE.

Phan-Minh, T.; Grigore, E. C.; Boulton, F. A.; Beijbom, O.; and Wolff, E. M. 2020. Covernet: Multimodal behavior prediction using trajectory sets. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 14074–14083.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.

Sadeghian, A.; Kosaraju, V.; Sadeghian, A.; Hirose, N.; Rezatofighi, H.; and Savarese, S. 2019. Sophie: An attentive gan for predicting paths compliant to social and physical constraints. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 1349–1358.

Salzmann, T.; Ivanovic, B.; Chakravarty, P.; and Pavone, M. 2020. Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XVIII 16*, 683–700. Springer.

Sha, L.; Lucey, P.; Zheng, S.; Kim, T.; Yue, Y.; and Sridharan, S. 2017. Fine-grained retrieval of sports plays using tree-based alignment of trajectories. *arXiv preprint arXiv:1710.02255*.

Sun, C.; Karlsson, P.; Wu, J.; Tenenbaum, J. B.; and Murphy, K. 2019. Predicting the present and future states of multi-agent systems from partially-observed visual data. In *International Conference on Learning Representations*.

Tuyls, K.; Omidshafiei, S.; Muller, P.; Wang, Z.; Connor, J.; Hennes, D.; Graham, I.; Spearman, W.; Waskett, T.; Steel, D.; et al. 2021. Game Plan: What AI can do for Football, and What Football can do for AI. *Journal of Artificial Intelligence Research*, 71: 41–88.

Van den Berg, J.; Lin, M.; and Manocha, D. 2008. Reciprocal velocity obstacles for real-time multi-agent navigation. In *2008 IEEE international conference on robotics and automation*, 1928–1935. IEEE.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Vidal-Codina, F.; Evans, N.; El Fakir, B.; and Billingham, J. 2022. Automatic event detection in football using tracking data. *Sports Engineering*, 25(1): 18.

Wang, J.; Hertzmann, A.; and Fleet, D. J. 2005. Gaussian process dynamical models. *Advances in neural information processing systems*, 18.

Wang, R.; Wang, S.; Yan, H.; and Wang, X. 2023. Wsip: Wave superposition inspired pooling for dynamic interactions-aware trajectory prediction. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, 4685–4692.

Wang, Z.; Veličković, P.; Hennes, D.; Tomašev, N.; Prince, L.; Kaisers, M.; Bachrach, Y.; Elie, R.; Wenliang, L. K.; Piccinini, F.; et al. 2024. TacticAI: an AI assistant for football tactics. *Nature communications*, 15(1): 1906.

Xia, T.; Qi, Y.; Feng, J.; Xu, F.; Sun, F.; Guo, D.; and Li, Y. 2021. Attnmove: History enhanced trajectory recovery via attentional network. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, 4494–4502.

Xu, Y.; Bazarjani, A.; Chi, H.-g.; Choi, C.; and Fu, Y. 2023. Uncovering the missing pattern: Unified framework towards trajectory imputation and prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 9632–9643.

Yeh, R. A.; Schwing, A. G.; Huang, J.; and Murphy, K. 2019. Diverse generation for multi-agent sports games. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4610–4619.

Yuan, Y.; Weng, X.; Ou, Y.; and Kitani, K. M. 2021. Agentformer: Agent-aware transformers for socio-temporal multi-agent forecasting. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 9813–9823.

Zhan, E.; Zheng, S.; Yue, Y.; and Lucey, P. 2018. Generative multi-agent behavioral cloning. *arXiv preprint arXiv:1803.07612*, 2.

Zhang, P.; Ouyang, W.; Zhang, P.; Xue, J.; and Zheng, N. 2019. Sr-lstm: State refinement for lstm towards pedestrian trajectory prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 12085–12094.

Zheng, S.; Yue, Y.; and Lucey, P. 2016. Generating long-term trajectories using deep hierarchical networks. *Advances in Neural Information Processing Systems*, 29.

Zhou, Z.; Ye, L.; Wang, J.; Wu, K.; and Lu, K. 2022. Hivt: Hierarchical vector transformer for multi-agent motion prediction. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 8823–8833.