arXiv:2610.05961v1 [cs.LG] 5 Oct 2026

Strategic Multi-Agent Learning for Interpretable Action Valuation of All Players in Football

Kenjiro Ide[1], Taiga Someya[2], Kohei Kawaguchi[3], Keisuke Fujii[1,4*]

[1]Graduate School of Informatics, Nagoya University, Nagoya, Japan.
[2]Graduate School of Arts and Sciences, The University of Tokyo, Tokyo, Japan.
[3]Department of Economics, The Hong Kong University of Science and Technology, Hong Kong, China.
[4]Center for Advanced Intelligence Project, RIKEN, Osaka, Japan.

*Corresponding author(s). E-mail(s): fujii@i.nagoya-u.ac.jp;
Contributing authors: ide.kenjiro@g.sp.m.is.nagoya-u.ac.jp;
tsomeya@g.ecc.u-tokyo.ac.jp; kkawaguchi@ust.hk;

Abstract

Valuing player actions in football requires accounting for continuous strategic interactions among 22 players, including off-ball movements and defensive positioning. Existing reinforcement-learning-based valuation methods commonly aggregate decisions at the team level or estimate player values independently, leaving strategic interdependence among players insufficiently represented. This study proposes an action valuation framework inspired by Markov perfect equilibrium (MPE) for on-ball, off-ball, and defensive players in football. The framework formulates each possession as a finite-horizon dynamic game and represents each player as an autonomous agent whose policy depends on the current game state. It uses MPE as a motivating solution concept but does not compute or verify an exact equilibrium strategy profile. To improve interpretability, we use Expandable Decision-Making States (EDMS) as a tactically interpretable state representation and decompose the Q-value into a successor-feature basis and a linear reward-weight vector. The value basis is estimated using a two-stage learning procedure that combines linear TD initialization with nonlinear iterative refinement. Using tracking and event data from 95 J1 League matches, we compare the proposed formulation with an independent reinforcement learning baseline. Because the two formulations define TD errors in different target spaces, TD MSE is used only to assess consistency within each formulation. With EDMS held fixed, the independent baseline assigns the highest value to forward movement in 99.21% of the

evaluated off-ball states, whereas the most frequent direction under the proposed formulation accounts for 17.63%. The association between team-level average Q-values and season-level expected goals is negative for the independent baseline ($\boldsymbol{\rho = -0.37}$) and weakly positive for the proposed formulation ($\boldsymbol{\rho = 0.29}$). Qualitative analyses further provide examples of context-dependent valuations of off-ball movements and defensive positioning. These results suggest that the proposed learning formulation as a whole produces more context-sensitive action rankings than the independent RL baseline in the evaluated data, although the current comparison does not isolate the contribution of the MPE-inspired component.

Keywords: Machine learning, Reinforcement learning, Multi-agent, Game theory, Football

1 Introduction

In invasion-type team sports such as football, the state space is inherently high-dimensional, as all players continuously interact on a two-dimensional field, and these dimensions are strongly coupled through player interactions. Early rule-based studies addressed this complexity by applying geometric or kinematic rules to estimate ball access and areas of control [1, 2]. While these models are intuitive for coaches, they are typically limited to narrowly defined situations such as shots, pressure, and crosses, making it difficult to extend them to evaluate all actions of on-ball, off-ball, and defensive players.

Subsequently, many studies in machine learning have evaluated performance by predicting discrete events [3, 4] or forecasting future trajectories [5]. Although these approaches are powerful for prediction and useful for many practical analytical tasks, they are primarily based on statistical mappings from observed states to future events or outcomes and do not explicitly model each player as an autonomous decision-making agent with rewards and policies.

More recently, deep reinforcement learning (DRL) has been adopted to directly estimate action-value functions from event and tracking data [6]. DRL enables the direct learning of value functions from raw data, capturing the long-term strategic impact of actions and allowing for alternative action evaluation beyond prediction. However, the interpretability of the learned value functions and policies remains a significant challenge. To address this issue, prior work [7] proposed Expandable Decision-Making States (EDMS). EDMS enhances interpretability by incorporating human-interpretable tactical concepts, such as marking pressure, passing lanes, and space control, into the state representation alongside raw positional data. Nevertheless, existing DRL-based approaches in football [6–9] model teams or players independently and do not capture the interactions among individual players' decision-making processes.

To address this limitation, this study incorporates a dynamic-game perspective [10] into a DRL-based action valuation framework and proposes a formulation inspired by Markov Perfect Equilibrium (MPE). MPE characterizes a Markov strategy profile

in which each player's policy is a best response to the policies of other players at every game state. We use this perspective to formulate each football possession as a finite-horizon dynamic game, with on-ball attackers, off-ball attackers, and defenders represented as autonomous decision-making agents. Rather than computing an exact equilibrium or explicitly enumerating all joint action profiles, the framework estimates player-specific value functions from observed multi-player states and transitions.

To make these action values interpretable, EDMS provides a tactically interpretable state representation that describes pressure, space, passing options, shooting opportunities, and possession context. The Q-value is then decomposed into successor-feature basis and a linear reward-weight vector, separating the estimated future accumulation of tactical reward features from the specification of tactical preferences. Finally, the successor-feature basis is estimated through a two-stage learning procedure consisting of linear TD initialization followed by nonlinear iterative refinement, allowing the model to capture nonlinear state-dependent patterns in the observed transitions.

The contributions of this study are summarized as follows:

1. We formulate football action valuation as a finite-horizon dynamic game and propose an MPE-inspired framework for estimating the action values of on-ball attackers, off-ball attackers, and defenders from observed multi-player transitions. Rather than computing an equilibrium strategy profile, the framework uses the MPE perspective to motivate state-dependent multi-agent value equations and estimates their continuation values from the observed transitions.
2. We combine an interpretable successor-feature representation with a two-stage estimation procedure, in which linear semi-gradient TD estimates are used to pretrain a nonlinear value approximator before iterative refinement. This representation separates the estimated accumulation of tactical reward features from their predefined weights and permits post-hoc analysis under alternative reward weights.
3. We compare independent RL and the proposed MPE-inspired formulation under both PVS and EDMS state representations. We evaluate within-formulation TD consistency, the concentration of off-ball action rankings, and the association between team-level Q-values and season-level xG, together with qualitative case studies of off-ball and defensive positioning.

2 Related work

2.1 Player performance evaluation in football

Research on play evaluation in football analytics aims to quantify the impact of individual actions on match outcomes, and a wide variety of approaches have been proposed. Early studies primarily focused on shots, leading to the widespread adoption of Expected Goals (xG) models [11], which estimate scoring probabilities based on shot location and contextual factors. However, since xG is limited to valuing shots, there has been a growing demand for more comprehensive evaluation frameworks.

One such approach evaluates all on-ball actions. VAEP (Valuing Actions by Estimating Probabilities) [3] proposes a framework that assigns value to actions such as

passes and dribbles based on how they change the probabilities of scoring and conceding. Similar approaches have also been applied to pass evaluation, where studies quantify both the probability of interception and the likelihood of advancing an attack [12]. By decomposing the sparse reward of goals into denser measures of contribution to scoring probability, these methods enable the unified evaluation of all actions on a common scale.

With the advent of tracking data, it has become possible to evaluate off-ball movements and spatial value. The Expected Possession Value (EPV) model [13] estimates the value of possession at any moment on the pitch using deep learning, thereby enabling contextual evaluation that includes off-ball dynamics. Similarly, Expected Threat (xT) [14, 15] introduces the concept of assigning threat values to different pitch areas, allowing the evaluation of actions that move the ball into more advantageous regions.

Beyond individual actions, recent studies have also sought to quantify team-level tactical behavior. In addition to classical approaches that evaluate spatial dominance using Voronoi diagrams formed by player positions [2, 16], more advanced methods incorporate player velocity and acceleration to model dynamic control regions [17]. Other studies measure how effectively players' movements create space for teammates [18]. Defensive analysis has also progressed, with frameworks such as GVDEP [19], which evaluates how defensive structures reduce the probability of conceding, and models that assess team defense based on probabilities of ball recovery and opponent attacks [20].

Recent advances in deep learning have further enabled approaches that treat play evaluation as a prediction task. Transformer-based models can predict what type of action will occur next [4], as well as when and where it will occur [21]. In addition to event prediction, trajectory prediction [22, 23] and its application to play evaluation [5, 24] have also been explored. Furthermore, tactical support systems such as TacticAI [25] have been developed to learn effective strategies in specific situations, such as corner kicks, and recommend optimal plays. While these predictive models are powerful for post-learning evaluation and future scenario forecasting, their analyses are primarily based on statistical mappings from observed states to future events or outcomes. In contrast, explicitly modeling each player as an autonomous decision-making agent with internal objectives and policies enables the capture of underlying tactical intent, allowing for a more comprehensive and multi-perspective evaluation of play quality.

2.2 Reinforcement learning in sports

Reinforcement learning (RL) has been applied not only to football but also to the analysis of various team and individual sports. In the forward (simulation-based) approach, research in football has been significantly accelerated by the introduction of the open-source simulation environment Google Research Football (GRF) [26]. GRF provides a standardized platform for training and valuing agents, and since its introduction, a wide range of methods have been developed to learn complex cooperative behaviors. However, it has been pointed out that a strategic gap exists between AI trained in simulation and the behavior of real-world players [27–29]. To bridge this gap,

more advanced learning approaches have been proposed, such as combining curriculum learning with self-play to achieve human-level tactical performance [30].

In contrast, the inverse (data-driven) approach leverages real-world data. Early work in ice hockey constructed Markov game models using play-by-play data to evaluate context-dependent action values of players [31]. Similarly, in racket sports such as badminton, deep reinforcement learning has been applied to player evaluation [32]. More recently, frameworks have been developed to support optimal decision-making based on real match data. For example, some studies focus on optimizing players' action selection to maximize expected possession value [8], while others employ inverse reinforcement learning to infer latent strategies underlying offensive and defensive plays [33, 34].

Among these studies, the work of Nakahara et al. [6] is particularly noteworthy, as it explicitly models individual players within a multi-agent reinforcement learning (MARL) framework using real tracking data, enabling the evaluation of both on-ball and off-ball actions.

While Nakahara et al. [6] provide a pioneering application of MARL to evaluate on-ball and off-ball players using real tracking data, their framework treats player interactions without an explicit equilibrium-based formulation. The novelty of our work lies in formulating football action valuation from an MPE-inspired multi-agent learning perspective. We adopt the previously proposed EDMS [9] as a tactically interpretable state representation within this framework. This combination enables the model to represent strategic interactions among players while retaining interpretability through football-specific state variables.

2.3 Game theory in sports

Sports, particularly competitive games such as football and basketball, are inherently settings of strategic interaction, where each player's actions influence the outcomes of others. As such, the application of game theory provides a natural analytical framework. Game-theoretic studies in sports can be broadly categorized into static and dynamic approaches.

Static game-theoretic approaches have primarily focused on isolated, one-shot events such as penalty kicks. PK situations can be readily modeled as zero-sum games between the kicker and the goalkeeper, and numerous studies have examined whether professional players behave according to mixed-strategy Nash equilibria [35]. Chiappori et al. [36] conducted statistical tests accounting for player heterogeneity and demonstrated that professional players largely conform to theoretical equilibrium strategies. In contrast, Bar-Eli and Azar [37] combined survey data from goalkeepers with video analysis to show that theoretically optimal strategies, such as aiming for the upper corners of the goal, are often distorted by psychological factors, including risk aversion and action bias, highlighting the limitations of purely rational models. Furthermore, cooperative game-theoretic approaches have also been explored, where player contributions to team performance are quantified using Shapley values and compared with conventional performance metrics [38].

Dynamic game-theoretic research, which considers the temporal evolution of play and state transitions, has aimed to model the full dynamics of matches. From a physics-inspired perspective, some studies model interactions between players as spring-like forces and describe collective team movement and spatiotemporal dynamics using stochastic differential equations [39]. Other approaches, such as Soccer-Decoder, simulate match progression and team performance decay using differential equations [40]. While these models are useful for reproducing macro-level match dynamics, they are often too coarse to capture individual players' decision-making processes.

More tactically oriented studies have focused on decision-making under uncertainty. Beal et al. [41] modeled pre-match tactical decisions as Bayesian games and in-game adjustments, such as substitutions and tactical changes, as stochastic games, constructing a framework to maximize win probability based on empirical data. Lennartsson et al. [42] defined a "potential field" of scoring probability and proposed a mathematical formulation of game intelligence in which defensive players act to minimize this potential. Additionally, data mining approaches have been proposed in simulated environments, where opponent strategies are learned as automata and counter-strategies are explored [43].

Despite these advances, traditional game-theoretic approaches face significant limitations. Static models ignore the continuity of play, while many dynamic models are constrained by computational complexity, forcing them to either drastically simplify the state space or focus on high-level "coach-centric" decision-making. In real matches, however, players make instantaneous, fine-grained decisions within a continuous spatiotemporal environment, based on a vast and complex state space defined by the positions and movements of teammates and opponents. Deriving Nash equilibria analytically in such environments is, in practice, intractable.

The MPE-inspired framework adopted in this study can be viewed as a learning-based approximation to dynamic-game value modeling in such computationally challenging settings. By combining this MPE-inspired formulation with EDMS, the proposed model represents tactically meaningful state variables that are often abstracted away in conventional game-theoretic models, enabling action values to be estimated from real tracking and event data.

3 Formulation

In this section, we formulate a football possession sequence as a finite-horizon dynamic game among multiple decision-making agents. Section 3.1 defines the state, action, transition, and reward setting. Section 3.2 introduces MPE as a motivating solution concept and clarifies how its associated value equations are approximated without computing an exact equilibrium strategy profile. Section 3.3 defines the tactical reward features and the successor-feature-based value representation. The corresponding two-stage estimation procedure is presented in Section 4.

3.1 Problem Setting

Each football possession sequence is formulated as a finite-horizon stochastic game involving a set of players $I = \{1, \ldots, N\}$. In the empirical implementation, $N = 22$,

corresponding to 11 attacking and 11 defending player slots. The index i denotes a focal-player slot within a possession rather than a persistent player identity across matches. Each possession contains time steps $t = 0, \ldots, T$, where the terminal time T may differ across possessions. At each time step t, the game is described by a state $s_t \in S$, which is constructed from tracking and event data. Each player i selects an action $a_{i,t} \in A_i(s_t)$, and the joint action is denoted by $a_t = (a_{1,t}, \ldots, a_{N,t})$. The next state is generated according to the transition probability $P(s_{t+1} \mid s_t, a_t)$, which depends on the actions of all players.

We consider two alternative state representations: Position and Velocity State (PVS), which describes the physical configuration of the players and ball, and Expandable Decision-Making State (EDMS), which represents tactically interpretable information. The definitions of PVS and EDMS follow [6, 9]. Table 1 summarizes the dimensionality and main variables of the two state representations used in the experiments.

Table 1 Summary of the state representations used in the experiments.

State representation	Dimension	Main variables
PVS	92	Player and ball positions and velocities.
EDMS	74	Pressure, space, passing options, shooting opportunities, offside-line information, and possession context.

Following previous studies [6, 9], we use a discrete action space consisting of 15 feasible action classes. As summarized in Table 2, the action space comprises five on-ball attacking actions, one aggregated defensive action, and nine off-ball actions. A separate padding label is used during sequence batching but is not treated as a feasible player action and is therefore excluded from the 15 action classes.

Table 2 Summary of the feasible discrete action classes.

Action category	No. of classes	Actions
On-ball attacking actions	5	Pass, through pass, shot, cross, and dribble
Defensive action	1	Tackle, interception, clearance, or block, aggregated into one defensive-action class
Off-ball actions	9	Movement in eight directions and idle
Total	15	

Because not every action is feasible for every player, the available action set is determined by the player's possession role at state s_t. Table 3 summarizes the corresponding state-dependent action masking. The attacking ball carrier is restricted to on-ball attacking actions, attacking off-ball players to the nine off-ball actions, and defending players to the off-ball actions and the aggregated defensive action.

Accordingly, each player has a state-dependent feasible action set $A_i(s_t)$. Each player receives a reward $r_i(s_t, a_t)$, whose feature-based definition is provided in Section 3.3. The corresponding Markov policies and finite-horizon value functions are introduced in Section 3.2.

Table 3 State-dependent action masking according to player role. Actions not listed as feasible are masked.

Player role	Feasible actions
Attacking ball carrier	Pass, through pass, shot, cross, and dribble
Attacking off-ball player	Movement in eight directions and idle
Defending player	Movement in eight directions, idle, and defensive action

3.2 Markov Perfect Equilibrium (MPE)

In this study, MPE serves as the motivating solution concept for modeling state-dependent strategic behavior in a dynamic multi-agent setting.

In dynamic multi-stage games, a Nash equilibrium may rely on continuation strategies that are not credible after histories that are not reached on the equilibrium path. Subgame-perfect equilibrium (SPE) strengthens Nash equilibrium by requiring the strategy profile to constitute a Nash equilibrium in every subgame, thereby ruling out such non-credible threats. However, a general SPE may allow strategies to depend on the full history of play. In environments such as football, where the state space is large and continuous, representing history-dependent strategies is computationally impractical. We therefore adopt an MPE perspective and restrict attention to Markov strategies, in which each player's strategy depends only on the current game state.

Accordingly, for each player $i \in I$, a Markov strategy π_i assigns a probability distribution over the state-dependent feasible action set:

$$\pi_i(\cdot \mid s_t) \in \Delta(A_i(s_t)). \tag{1}$$

An MPE is a strategy profile $\pi^* = (\pi_1^*, \ldots, \pi_N^*)$ such that no player can improve its expected return by unilaterally changing its strategy while the strategies of the other players remain fixed. Formally, for every player i, state $s_t \in S$, and alternative Markov strategy $\tilde{\pi}_i$, the following condition holds:

$$V_i^{\pi^*}(s_t) \geq V_i^{(\tilde{\pi}_i, \pi^*_{-i})}(s_t), \qquad \forall \tilde{\pi}_i,\ \forall s_t \in S. \tag{2}$$

Here, V_i^{π} denotes the value function for player i at state s_t under the strategy profile π:

$$V_i^{\pi}(s_t) = \mathbb{E}_{\pi}\left[\sum_{k=t}^{T} \gamma^{k-t} r_i(s_k, a_k) \,\middle|\, s_t\right]. \tag{3}$$

In this study, the discount factor is set to $\gamma = 1$. Because each possession has a finite terminal time T, the resulting return remains well defined without temporal discounting. This setting allows rewards obtained near the end of a possession to be propagated to earlier actions without temporal attenuation, including actions involved in build-up play.

From the perspective of player i, fixing the Markov strategies of the other players induces a state-dependent stochastic environment in which their action distributions depend on the current state s_t. The resulting decision problem can therefore be interpreted as an induced Markov decision process for player i. The MPE condition requires the state-dependent strategies of all players to be mutually consistent best responses.

The empirical implementation does not explicitly estimate the Markov policies π_i or verify an equilibrium strategy profile. Instead, observed player actions are used to construct Bellman-style continuation targets from the state-action transitions in the dataset. Specifically, continuation values are bootstrapped using the next actions observed in each possession rather than actions obtained from an explicit best-response or greedy-policy update. The effects of the other players are represented through the observed multi-player state and realized transitions, avoiding explicit enumeration of all joint action profiles. The resulting estimation procedure is therefore MPE-inspired but does not constitute an algorithm for computing an MPE.

3.3 Interpretable Value Representation and Reward Design

To make the MPE-inspired action valuation framework interpretable, we do not learn a black-box scalar Q-value directly. Instead, we decompose the Q-value into a value basis and a reward weight vector. The value basis represents the expected future accumulation of tactically meaningful reward features, whereas the reward weight vector specifies how these features are valued. This decomposition separates the estimated future accumulation of tactical reward features from the weights used to combine them.

For this purpose, the immediate reward of player i is defined as a linear combination of reward features:

$$r_i(s_t, a_t) = z_i(s_t, a_t)^\top \theta, \tag{4}$$

where $z_i(s_t, a_t)$ denotes the reward-feature vector and θ is the reward-weight vector. The vector $z_i(s_t, a_t)$ is player-specific for attacking players, whereas it is shared among defending players to reflect their common objective of suppressing the opponent's attacking value.

In this study, we introduce $d = 3$ features corresponding to shot value, possession termination, and attacking-third ball occupancy. Let $V(x_t, y_t)$ denote the Expected Possession Value (EPV) [1] obtained from the EPV grid at the ball location (x_t, y_t) at time t. Let $S_{i,t} = \mathbf{1}\{a_{i,t} = \text{shot}\}$ and let E_t indicate the final time step of a possession segment. Let L denote the pitch length, with the origin located at the center of the pitch and the attacking direction aligned with the positive x-axis. In our setting, $L = 105$ m, so the attacking-third boundary is $x = L/6 = 17.5$ m. Let $\mathcal{I}_{\text{att}}$ denote the set of attacking players. We define

$$C_t = E_t \mathbf{1}\left\{ \sum_{j \in \mathcal{I}_{\text{att}}} S_{j,t} = 0 \right\} \mathbf{1}\{x_t < L/6\}. \tag{5}$$

For an attacking player i, the reward-feature vector is defined as

$$z_i(s_t, a_t) = \begin{bmatrix} v_t S_{i,t} \\ -V(-x_t, y_t) C_t \\ \mathbf{1}\{x_t \geq L/6\} \end{bmatrix}. \tag{6}$$

For a defending player i, it is defined as

[1] https://github.com/Friends-of-Tracking-Data-FoTD/LaurieOnTracking

$$z_i(s_t, a_t) = \begin{bmatrix} -\mathbf{1}\left\{\sum_{j \in \mathcal{I}_{\text{att}}} v_t S_{j,t} > 0\right\} \\ V(-x_t, y_t) C_t \\ 0 \end{bmatrix}. \tag{7}$$

Here, v_t denotes the configured shot value. The three components have the following interpretations:

- $z_{i,1}$ (Shot value): Nonzero only at a shot time. The shooting attacker receives v_t (e.g., $V(x_t, y_t)$ under the EPV specification), while all defenders receive -1.
- $z_{i,2}$ (Possession termination adjustment): Nonzero only at the final time step of a possession that ends without a shot and outside the attacking third. Its magnitude is the opponent-perspective EPV, $V(-x_t, y_t)$, with opposite signs for attackers and defenders.
- $z_{i,3}$ (Attacking-third ball occupancy): Equals one for every attacking player whenever the ball is located in the attacking third, $x_t \geq L/6$, and zero otherwise. Thus, it represents ball occupancy of the attacking third rather than only the first entry into it.

The weight parameters are predefined as $\theta = [1.0, -0.5, 0.2]^\top$. These weights were set heuristically to reflect the relative importance of the three reward components: shot value, possession termination adjustment, and attacking-third ball occupancy. They were not estimated from data in this study.

Based on the linearity of the reward function, the action-value function (Q-function) can also be decomposed into the inner product of the expected discounted sum of features and the weight vector. Given a Markov strategy profile, the action-value function $Q_i^\pi(s_t, a_{i,t})$ for agent i following policy π_i is given by:

$$Q_i^\pi(s_t, a_{i,t}) = h_i^\pi(s_t, a_{i,t})^\top \theta, \tag{8}$$

where $h_i^\pi(s_t, a_{i,t})$, referred to as the value basis or successor features, represents the expected discounted sum of the feature vector z_i obtained by taking action $a_{i,t}$ at state s_t and thereafter following policy π (with other agents' policies π^{-i} treated as part of the environment):

$$h_i^\pi(s_t, a_{i,t}) = \mathbb{E}^\pi \left[\sum_{k=t}^{T} \gamma^{k-t} z_i(s_k, \mathbf{a}_k) \mid s_t, a_{i,t} \right]. \tag{9}$$

Here, the expectation is also taken over the current actions of the other players, $a_{-i,t} \sim \pi_{-i}(\cdot \mid s_t)$. In the empirical TD updates, this expectation is approximated using the realized joint actions in the observed state-action transition data sequences, rather than by explicitly enumerating all possible joint action profiles.

The advantage of this linear decomposition is that it separates the learning of the value function into two components: the learned value basis h_i^π, which represents the expected future accumulation of tactical reward features, and the reward structure, which is specified by θ. This improves interpretability by revealing which tactical components contribute to the estimated Q-value. Once the value basis has been estimated, alternative reward weights can be applied post hoc to inspect how tactical preferences

change local action rankings. In the current implementation, the bootstrap targets are constructed using the next actions observed in the possession sequences rather than greedy actions selected according to the reward weights. Therefore, this reweighting changes the valuation of the fixed learned value basis without re-estimating the underlying continuation values.

4 Method

4.1 Learning Objective

Our estimation procedure adapts the temporal-difference framework of Adusumilli and Eckardt [10], who propose linear semi-gradient and approximate value-iteration estimators for recursive value terms in dynamic discrete choice models and dynamic games. Their primary objective is structural parameter estimation, where the recursive value terms are subsequently used in a conditional-choice-probability-based pseudo-likelihood. In contrast, our objective is football action valuation: we fix the tactical reward weights θ and directly estimate the successor-feature basis h_i. We further combine the two estimation approaches sequentially: the linear TD estimates are used as supervision for pretraining a nonlinear value approximator, which is subsequently refined by AVI, rather than treating the two estimators as alternatives. In our implementation, both stages estimate continuation values using the next actions observed in the possession sequences rather than actions selected by a greedy or explicit best-response update. These adaptations allow the TD framework to be applied to finite-horizon football possessions with nonlinear value approximation from observed state-action transition data.

The objective of the learning procedure is not to compute an exact MPE. Instead, it estimates successor-feature value bases that approximate Bellman-style value equations motivated by MPE-inspired dynamic-game formulation. The previous section defined the action valuation framework and its interpretable value representation. This section describes how the value basis is estimated from real tracking and event data.

We adopt a two-stage learning approach. First, a linear semi-gradient TD estimator produces initial estimates of the successor-feature basis on the observed state-action pairs. A randomly initialized MLP is then pretrained to reproduce these linear TD estimates by minimizing their squared prediction error; thus, the linear coefficients are not directly transferred to the network weights. Second, the pretrained MLP is refined by approximate value iteration using Bellman targets constructed from the observed next state-action pairs, allowing nonlinear state-dependent patterns to be captured.

4.2 Linear TD Initialization of Successor Features

In the first stage of learning, the value basis $h_i(s_t, a_{i,t}) \in \mathbb{R}^d$ is approximated using a linear model:

$$h_i(s_t, a_{i,t}; W_i) = W_i^\top \phi_i(s_t, a_{i,t}), \tag{10}$$

where $\phi_i(s_t, a_{i,t}) \in \mathbb{R}^{p_i}$ denotes the basis-function vector for focal-player slot i, and $W_i \in \mathbb{R}^{p_i \times d}$ is the parameter matrix to be learned. A separate basis pipeline is constructed for each of the 22 focal-player slots, corresponding to the 11 attacking and 11 defending positions in each frame rather than to persistent player identities. Let $o_{i,t}$ denote the player-centric state representation for focal player i, which contains its relative tactical context, global pitch context, and focal-player location.

For each focal-player slot, the observed action is represented by its integer action ID, and the value basis contains a separate action block for every action category. Continuous components of $o_{i,t}$ are standardized using the mean and standard deviation estimated from the corresponding training observations. The basis vector is then constructed as

$$\tilde{o}_{i,t} = \text{Standardize}_i(o_{i,t}), \tag{11}$$

Let $c_{i,t}$ denote the role category of focal player i at time t.

$$\phi_i(s_t, a_{i,t}) = \mathbf{e}_{(c_{i,t}, a_{i,t})} \otimes \begin{bmatrix} 1 \\ \tilde{o}_{i,t} \end{bmatrix}, \tag{12}$$

where $\mathbf{e}_{(c_{i,t}, a_{i,t})}$ is a one-hot vector identifying the corresponding role–action combination. This block-structured representation allows the coefficients of the state variables to vary across role-action combinations. It does not include higher-order state-state polynomial terms.

In this study, $d = 3$, corresponding to the three reward features defined in Section 3.3. Each column of W_i therefore estimates the expected discounted accumulation of one reward feature.

To estimate W_i, we use a stochastic semi-gradient TD update based on observed transition data $(s_t, a_t, z_i(s_t, a_t), s_{t+1}, a_{i,t+1})$, where $a_{i,t+1}$ denotes the next action of player i observed in the data. The objective is

$$L(W_i) = \mathbb{E}\left[\|z_i(s_t, a_t) + \gamma h_i(s_{t+1}, a_{i,t+1}; W_i) - h_i(s_t, a_{i,t}; W_i)\|_2^2\right]. \tag{13}$$

Here, the TD target is defined as

$$Y_{i,t} = z_i(s_t, a_t) + \gamma h_i(s_{t+1}, a_{i,t+1}; W_i) \tag{14}$$

where $(s_{t+1}, a_{i,t+1})$ is the next observed state-action pair. At the final time step of a possession, the continuation value is set to zero, and the target is therefore

$$Y_{i,T} = z_i(s_T, a_T). \tag{15}$$

For possessions ending with a shot, the shot-value component $(z_{i,1})$ is thus assigned directly on the terminal row. For possessions ending without a shot, the possession-termination component $(z_{i,2})$ is assigned on the terminal row and enters the Bellman target directly. The attacking-third ball occupancy component $(z_{i,3})$ is assigned at every time step in which the ball is in the attacking third, including the terminal time step when applicable. These immediate reward features are propagated to preceding states through the estimated continuation values.

The parameter matrix is updated as

$$W_i \leftarrow W_i + \alpha\phi(s_t, a_{i,t}) \left(Y_{i,t} - h_i(s_t, a_{i,t}; W_i)\right)^\top, \tag{16}$$

where α denotes the learning rate. The resulting linear TD predictions are used as supervision for pretraining the randomly initialized MLP.

4.3 Approximate Value Iteration (AVI) for Nonlinear Value-Basis Refinement

In the second stage, we refine the value basis using a nonlinear function approximator implemented as a multilayer perceptron (MLP). The linear TD estimate obtained in the first stage is used as supervision for pretraining.

Unlike a greedy control update, the nonlinear refinement uses the next action actually observed in the transition data when constructing the continuation target. Thus, the procedure does not explicitly perform best-response updates or compute an equilibrium strategy profile. Instead, it iteratively fits the successor-feature value basis to Bellman-style targets constructed from observed multi-player transitions. The learning procedure is as follows.

1. **Pretraining:** The parameters of $h_i(s, a; \psi)$ are randomly initialized and then pretrained to reproduce the linear TD estimates obtained in the first stage by minimizing the squared prediction error. The resulting pretrained parameters are used as the initial parameters for the subsequent AVI refinement.
2. **Iteration** $(k = 0, 1, 2, \ldots)$**:**
   For each observed transition $(s_t, a_t, z_i(s_t, a_t), s_{t+1}, a^{\mathrm{obs}}_{i,t+1})$, construct the target

$$y^{(k)}_{i,t} = z_i(s_t, a_t) + \gamma h_i(s_{t+1}, a^{\mathrm{obs}}_{i,t+1}; \psi_k). \tag{17}$$

   The parameters are then updated by minimizing

$$\psi_{k+1} \leftarrow \arg\min_{\psi} \mathbb{E}_{\mathcal{D}} \left[\left\| y^{(k)}_{i,t} - h_i(s_t, a_{i,t}; \psi) \right\|^2 \right]. \tag{18}$$

3. **Convergence:** Repeat the update until the predefined convergence criterion is satisfied.

The resulting nonlinear value basis is subsequently used to compute action values for all feasible actions through $Q_i(s, a) = h_i(s, a)^\top \theta$. Greedy actions are derived only at this post-training inference stage for action-ranking analysis and are not fed back into the TD targets.

5 Experimental Results

5.1 Experimental Setup

We evaluated the proposed model using tracking and event data from 95 matches in the Meiji Yasuda J1 League during the 2023 and 2024 seasons, provided by Data Stadium Inc. The raw tracking data contain the positions of all players and the ball

recorded at 25 fps, and the event data include action annotations such as passes, shots, tackles, clearances, and interceptions. During preprocessing, the tracking data were resampled to 10 fps for model estimation. The dataset was split at the match level, using 50 matches from the 2023 season for training and 45 matches from the 2024 season for validation and testing, consisting of 5 validation and 40 test matches.

The tracking data were preprocessed by interpolating missing values and computing the velocity and acceleration of each player and the ball. The tracking and event data were then synchronized using timestamps. Following previous reinforcement learning studies on football action valuation, the attacking direction was normalized so that the attacking team always moves from left to right. Each match was segmented into attacking sequences based on the possession sequence identifier, and sequences with lengths between 30 and 300 time steps at 10 fps were retained.

After preprocessing and sequence extraction, the dataset contained 10,308 possession sequences and 1,561,255 event-sequence time steps. Independently, the total duration of the extracted possession sequences, calculated from the difference between each sequence's start and end frames, was 154,113.3 s (42.81 h). Specifically, the training, validation, and test sets contained 4,589, 918, and 4,801 possession sequences, respectively, comprising 684,605, 149,629, and 727,021 event-sequence time steps. Their total durations were 67,565.0 s (18.77 h), 14,783.6 s (4.11 h), and 71,764.7 s (19.93 h), respectively.

To examine whether the results depend on the state representation, we evaluated both the independent RL baseline and the MPE-inspired formulation using PVS and EDMS, resulting in four configurations: Independent+PVS [6], Independent+EDMS [9], MPE+PVS, and MPE+EDMS.

The independent baseline estimates each player's value function separately, whereas the proposed model estimates player-specific value bases within an MPE-inspired multi-agent formulation using a two-stage procedure consisting of linear TD initialization and nonlinear iterative TD refinement. Here, "player-specific" refers to conditioning the value estimate on a focal-player representation, rather than assigning a separate network to each player identity. At each time step, a focal-player representation is constructed for each of the 22 player slots, encoding the focal player's action and relative tactical context, including whether the player is the ball holder, an attacking off-ball player, or a defending player. Observations from all player slots are pooled during nonlinear AVI, and a shared MLP value approximator is used to estimate the successor-feature basis.

The independent baseline uses a single-layer GRU sequence encoder with an input size of 64 and a hidden size of 64. The baseline is trained using the Adam optimizer with a learning rate of 0.001 and weight decay of 10^{-5}. The discount factor is set to 1.0, the regularization weight to 10^{-5}, and the supervised action-loss weight to 0.005.

For the proposed MPE-inspired formulation, the h-value function was approximated by a nonlinear MLP with hidden-layer sizes of 64 and 32. Before nonlinear AVI refinement, a linear semi-gradient TD estimator was trained for one epoch with a learning rate of 0.001, and the resulting estimate was used as supervision for pretraining. The nonlinear refinement stage was run for at most 20 AVI iterations, with the value predictor trained for two epochs at each iteration. To stabilize the iterative

AVI updates, the predictor parameters from the previous AVI iteration were reused to initialize the next iteration, rather than reinitializing the network from scratch. A step-wise learning-rate schedule was also applied during nonlinear refinement, reducing the learning rate by a factor of 0.3 at AVI iterations 5, 10, and 15, with a minimum learning rate of 10^{-5}. A batch size of 2,048 was used for both the semi-gradient initialization and the nonlinear value-prediction updates. The validation attacker TD MSE was used as the primary criterion for early stopping and best-model selection, with patience-based stopping after at least five AVI iterations.

5.2 Performance Metrics

We used three quantitative metrics. First, the mean squared temporal-difference error (TD MSE) was used to assess internal consistency with the corresponding fitted TD targets within each learning formulation. For the MPE-inspired formulation, the TD residual is a three-dimensional vector corresponding to the three successor-feature components. For each focal-player slot, the squared residuals are summed across the three components and then averaged over samples. The reported TD MSE is obtained by taking the mean of these slot-level TD MSE values across the 22 focal-player slots. Because the independent RL baseline computes TD error in a scalar Q-value space, whereas the MPE-inspired formulation computes TD error in a successor-feature vector space, the absolute TD MSE values are not directly comparable across the two formulations. Second, the distribution of the highest-valued off-ball actions was analyzed as a diagnostic for action-ranking collapse. Although a more diverse distribution does not by itself prove tactical optimality, it indicates whether the learned action ranking avoids convergence to a trivial dominant movement. Third, as an exploratory team-level diagnostic, we examined the association between each team's Q-value summary and its official 2024 J1 League full-season expected goals (xG). Because the Q-value summaries are computed from the held-out tracking-data sample, whereas xG is a full-season club-level aggregate, this analysis is interpreted as a descriptive association with season-long attacking quality rather than as confirmatory evidence of external validity. We report Spearman's rank correlation coefficient ρ.

5.3 Quantitative Results

5.3.1 TD Consistency within Each Learning Formulation

We evaluated TD consistency for both the independent RL baseline [6, 9] and the proposed MPE-inspired formulation using PVS and EDMS. Table 4 reports the TD MSE for each configuration. Because the two learning formulations define TD errors in different target spaces, their absolute TD MSE values are not directly comparable.

Within the independent RL formulation, EDMS yielded a lower TD MSE than PVS. The same pattern was observed within the MPE-inspired formulation, where EDMS also yielded a lower TD MSE than PVS. These within-formulation comparisons indicate that the EDMS state representation provides better consistency with the corresponding TD targets under both learning procedures.

The absolute TD MSE values should not be compared between the two learning formulations. The independent RL baseline measures TD error in a scalar Q-value

Table 4 TD MSE for the independent RL and MPE-inspired formulations using PVS and EDMS. The independent RL TD error is defined over a scalar Q-value, whereas the MPE-inspired TD error is defined over the successor-feature vector. Therefore, TD MSE values should be compared only within the same learning formulation, not across formulations.

Learning formulation	TD target space	State representation	TD MSE
Independent RL [6]	Scalar Q-value	PVS	6.6947×10^{-4}
Independent RL [9]	Scalar Q-value	EDMS	4.5057×10^{-4}
MPE-inspired RL	Successor features ($d = 3$)	PVS	1.1229×10^{-2}
MPE-inspired RL	Successor features ($d = 3$)	EDMS	1.0876×10^{-2}

space, whereas the proposed method measures error in the $d = 3$ successor-feature space. Consequently, the two metrics differ in both target representation and scale. TD MSE is therefore used here only as a measure of internal consistency within each learning formulation and does not directly assess the tactical plausibility of alternative-action rankings.

5.3.2 Off-Ball Action Ranking Analysis

To complement the TD-based evaluation, we next examine the distribution of highest-valued off-ball actions. For this analysis, we compare the Independent RL + EDMS and MPE-inspired RL + EDMS configurations so that the state representation is held fixed.

Figure 1 provides a diagnostic view of the learned off-ball action rankings. The independent RL baseline assigns the highest Q-value to forward movement in 99.21% of the evaluated off-ball states. Thus, the model exhibits a highly concentrated off-ball action ranking. In contrast, the proposed MPE-inspired model produces a substantially more diverse distribution of highest-valued off-ball actions, with the most frequent direction accounting for only 17.63% of the evaluated states. Because both models use the same EDMS state representation in this comparison, the difference in action concentration cannot be attributed to the state representation alone. However, the two approaches also differ in model architecture, reward representation, learning objective, and training procedure. Therefore, the observed difference should be interpreted as a difference between the two learning formulations as a whole, rather than as evidence isolating the effect of the MPE-inspired component.

5.3.3 Exploratory Team-Level Association with Season-Level xG

Finally, we examine the association between the learned Q-values and season-level attacking performance as an exploratory team-level diagnostic.

Figure 2 shows the relationship between each team's Q-value summary computed from the 40 held-out test matches and official 2024 J1 League full-season xG. The xG values were obtained from the official J.League club statistics. [2] For each team, the Q-value summary was computed by averaging the Q-values of the observed actions over all evaluated player-time-step observations in the held-out test matches involving that team. Thus, the Q-value summaries and xG values differ in temporal coverage, with the former based on the held-out tracking-data sample and the latter covering

[2] https://www.jleague.jp/j1/stats/player/2024/expected_goals/search-list/

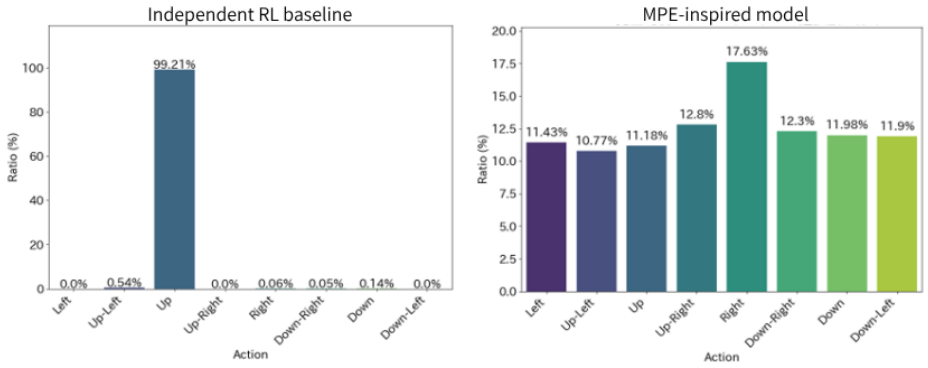


Fig. 1 Distribution of highest-valued off-ball actions obtained from post-training Q-value inference. The off-ball action space consists of eight directional movements and an idle action. The idle action was included when determining the highest-valued action but is omitted from the figure for visualization. The independent RL baseline assigns the highest Q-value to forward movement in almost all evaluated off-ball states, whereas the proposed MPE-inspired model distributes the highest-valued actions across multiple directions. "Up" corresponds to the attacking direction toward the opponent's goal.

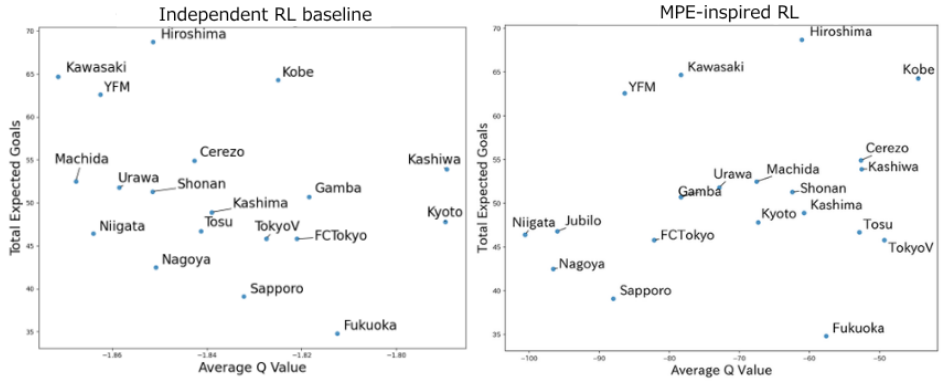


Fig. 2 Exploratory association between team-level Q-value summaries computed from the held-out tracking-data sample and official 2024 J1 League full-season xG. Each point represents one of the $n = 20$ matched teams. The independent RL baseline shows a negative Spearman rank correlation, whereas the proposed MPE-inspired model shows a weakly positive correlation.

the full 2024 season. Across the $n = 20$ matched teams, the independent RL baseline shows a negative correlation ($\rho = -0.37$), whereas the proposed MPE-inspired model shows a weakly positive association ($\rho = 0.29$).

These results indicate a difference in the direction of the team-level association between the two learning formulations, rather than demonstrating external validity. Given the small team-level sample and the differing temporal coverage of the two quantities, the results should be interpreted as descriptive. The modest magnitude of the associations may also reflect the fact that xG primarily reflects shot quality, whereas the proposed Q-values represent broader decision-making processes, including build-up play, off-ball movement, and defensive positioning.

Overall, these results show that the three diagnostics capture different aspects of the learned value functions. TD MSE measures consistency with the corresponding TD targets within each learning formulation, whereas the off-ball action distribution evaluates the behavior of alternative-action rankings and the xG analysis provides an exploratory team-level diagnostic of association with season-long attacking performance. Because these diagnostics measure different properties and the TD MSE target spaces differ between the two learning formulations, they should be interpreted separately rather than as directly comparable measures of model quality.

5.4 Qualitative Analysis

The quantitative results suggest that the proposed model avoids the highly concentrated forward-movement action ranking observed in the independent reinforcement learning baseline. To examine whether this difference corresponds to tactically meaningful decision-making, we analyze representative match situations and visualize the Q-values assigned to possible player movements. The examples were selected to illustrate qualitatively distinct tactical contexts, including open-space runs, support movements, scoring opportunities, and defensive positioning. The following analyses focus on off-ball behavior, defensive positioning, and the effect of reward-weight adjustment.

5.4.1 Evaluation of Off-ball Behavior

We first analyze whether the proposed model assigns tactically meaningful Q-values to off-ball movements in actual match situations. The following examples examine whether the model evaluates off-ball actions according to the surrounding tactical context, rather than simply assigning high values to forward movement.

Figure 3 shows a situation in which the target off-ball player has open space in front of him and the nearby defenders are not positioned to immediately close that space. In this case, the proposed model assigns relatively high Q-values to forward-oriented movements. This valuation is consistent with common tactical principles because moving into the open space would increase the receiver's separation from defenders and create a more progressive passing option for the ball carrier. This pattern is consistent with sensitivity to the surrounding defensive pressure.

Figure 4 shows a situation in which a nearby teammate has already moved into the forward space. Unlike the previous example, the target player's highest-valued movements shift toward a supporting position rather than direct forward movement. This pattern is consistent with the model accounting for the spatial occupation of teammates, as movements that would congest the same attacking lane receive lower values. The highest-valued support movement could provide a safer passing option for the ball carrier and maintain the continuity of the possession.

Figure 5 presents a situation near the opponent's goal, where small positional changes can directly affect the quality of the attacking opportunity. The proposed model assigns high Q-values to movements that improve the target player's relationship with the ball, the goal, and surrounding defenders. In particular, the model favors

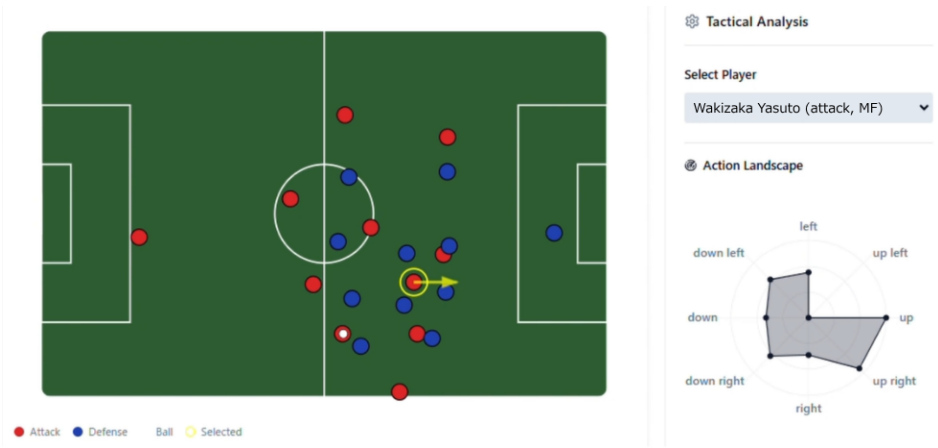


Fig. 3 Off-ball Q-value visualization. The left panel shows all players and the ball, with the white marker indicating the ball and the yellow-highlighted player indicating the target player whose Q-values are visualized. The right panel shows the target player's name and a radar chart of Q-values for off-ball movements in eight directions. "Up" corresponds to the attacking direction toward the opponent's goal. This example shows off-ball action valuation when open space is available ahead of the target player.

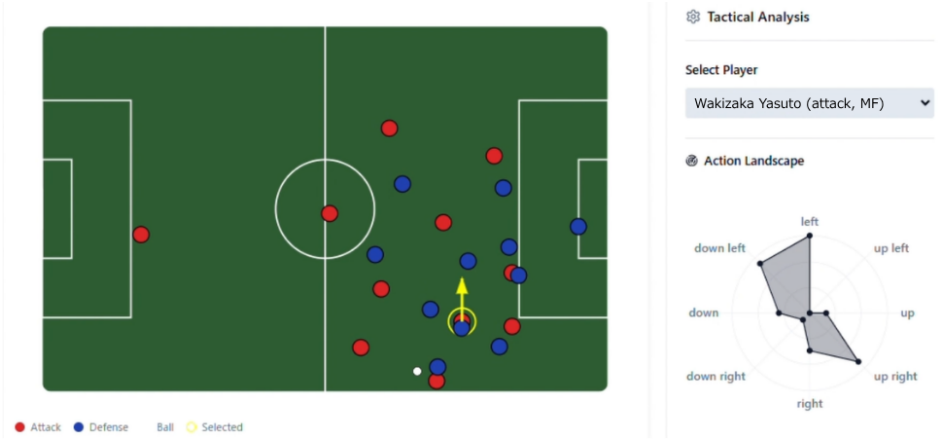


Fig. 4 Example of off-ball action valuation in response to a teammate's forward movement. The visualization layout follows Fig. 3.

movements that keep the player available as a passing option while avoiding immediate defensive pressure. This suggests that the learned Q-values reflect local attacking geometry, rather than only global indicators such as distance to goal.

These examples illustrate that the proposed model can assign Q-values that reflect contextual off-ball decision-making. Unlike the independent reinforcement learning baseline, whose highest-valued off-ball actions are strongly concentrated on forward movement, the proposed model evaluates multiple movement directions according to the surrounding tactical situation. These case studies support the interpretation

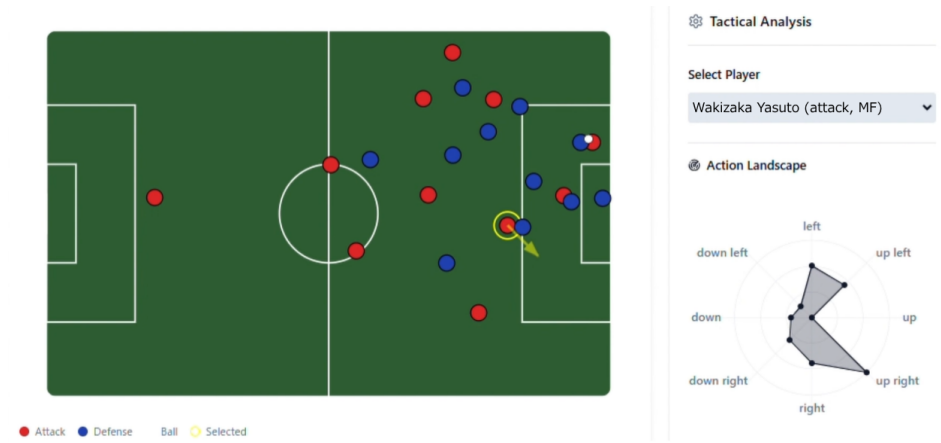


Fig. 5 Example of off-ball action valuation during a scoring opportunity. The visualization layout follows Fig. 3.

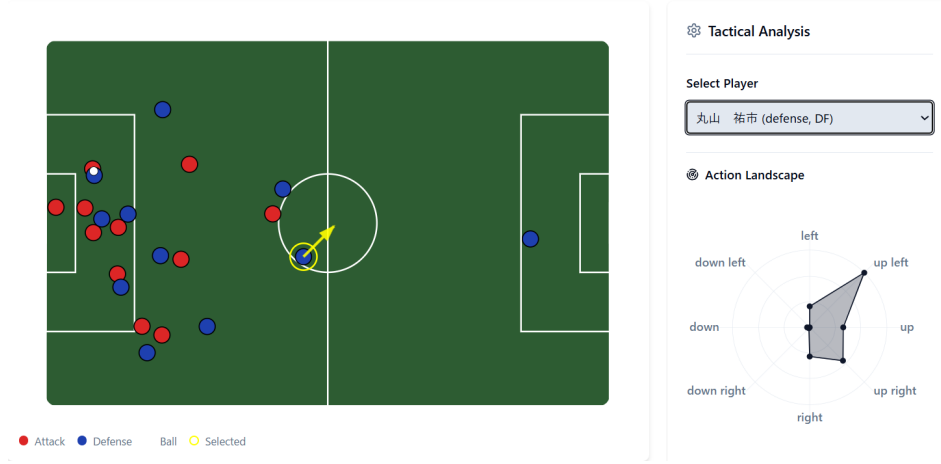


Fig. 6 Example of defensive positioning evaluation. The visualization layout follows Fig. 3.

that the proposed learning formulation produces more context-sensitive off-ball action rankings rather than consistently assigning the highest value to forward movement.

5.4.2 Evaluation of Defensive Positioning

A further advantage of the proposed framework is that it evaluates on-ball, off-ball, and defensive players. In contrast to valuation models [6] that are primarily evaluated through attacking outcomes, the proposed MPE-inspired formulation treats both attacking and defending players as decision-making agents. This enables the model to evaluate defensive movements that reduce the opponent's future attacking value.

Figure 6 shows an example of defensive positioning evaluation. The target defender is assigned high Q-values for movements that close dangerous space or block a potential



Fig. 7 Effect of reward weight adjustment on the highest-valued action. The visualization format follows Fig. 3. The lower-right panel shows the reward-feature weights θ. Under the default reward weights, $\theta = (1.0, -0.5, 0.2)^\top$, the highest-valued action is backward-right.

passing lane. This result suggests that the model can quantify defensive contributions that are difficult to capture using conventional on-ball event statistics. In particular, the proposed model can evaluate preventive defensive actions, such as occupying space before a pass is played or shifting position to reduce the opponent's future attacking options.

These results illustrate that the proposed framework can also assign context-dependent values to defensive positioning. In particular, movements that close dangerous space or restrict potential passing options can receive high Q-values even when they do not involve direct interaction with the ball. This suggests that the proposed framework can represent preventive defensive contributions that are difficult to capture using conventional on-ball action statistics.

5.4.3 Reflecting Tactical Preferences through Reward Weight Adjustment

This analysis is intended to demonstrate the interpretability and flexibility of the value representation rather than to introduce a separate model component. Because the Q-value is computed as the inner product between the learned value basis and the reward weight vector, different tactical preferences can be inspected after learning by reweighting the fixed value basis. This analysis should be interpreted as a post-hoc sensitivity analysis of local action rankings, rather than as a fully re-estimated value model under the new reward weights.

Figure 7 and Figure 8 compare the highest-valued actions under two different reward-weight settings. Figure 7 shows the result under the default setting, $\theta = (1.0, -0.5, 0.2)^\top$. Under this setting, the highest-valued action is backward-right. In



Fig. 8 Effect of increasing the attacking-third weight. The visualization format is the same as in Fig. 7. After changing the reward weights to $\theta = (1.0, -0.5, 1.0)^\top$, the highest-valued action changes from backward-right to forward-right.

contrast, Figure 8 shows the result after increasing the weight for attacking-third ball occupancy from 0.2 to 1.0. With this adjustment, the highest-valued action changes toward a more aggressive forward movement.

This comparison shows that the proposed model can incorporate tactical preferences in an interpretable manner. For example, if a coach or analyst wants to emphasize advancing the ball into the attacking third, this preference can be directly reflected by increasing the corresponding reward weight. Since this change is applied only to the reward weight vector, it can be performed without retraining the entire model. In the current implementation, the learned value basis is kept fixed during this analysis, and the reward weights are changed only when computing $Q_i(s,a) = h_i(s,a)^\top \theta$. Therefore, the analysis directly measures how alternative tactical reward preferences change the ranking of actions under the fixed learned value basis. In summary, the proposed framework functions not only as an action valuation model, but also as a tool for analyzing how different tactical priorities change the valuation of player decisions.

6 Conclusion

In this study, we proposed an MPE-inspired action valuation framework for on-ball, off-ball, and defensive players in football. The central idea of the framework is to formulate football as a dynamic game in which each player is modeled as an autonomous decision-making agent, and each player's action value is estimated within a Markovian multi-agent context that accounts for surrounding players through the observed game state and fitted value updates. This formulation allows the model to evaluate attacking, off-ball, and defensive actions while representing strategic context among players.

The proposed framework is implemented through three key components. First, EDMS provides a tactically interpretable state representation that encodes pressure, space, passing options, shooting opportunities, and possession context. Second, the Q-value is represented using successor features and a linear reward function, which enables interpretable decomposition of action values and post-hoc adjustment of tactical reward preferences. Third, a two-stage estimation procedure is used in which linear TD estimates are used to pretrain a nonlinear value approximator, which is subsequently refined through approximate value iteration to accommodate nonlinear state-dependent patterns in observed state-action transition data.

Using tracking and event data from 95 matches in the Meiji Yasuda J1 League, we compared the proposed learning formulation with an independent reinforcement learning baseline. Because the two formulations define TD errors in different target spaces, their absolute TD MSE values are not directly comparable; TD MSE was therefore used only to assess consistency within each formulation. In the EDMS-based comparison, the independent baseline assigned the highest value to forward movement in almost all evaluated off-ball states, whereas the proposed formulation produced a substantially more diverse distribution of highest-valued off-ball actions. Moreover, the correlation between team-level average Q-values and season-level xG changed from negative in the baseline to weakly positive in the proposed model, although the magnitude of this relationship remained limited.

The qualitative analyses further illustrated that the proposed model assigns context-dependent values to off-ball and defensive movements. The model evaluated movements into open space, support actions for the ball carrier, positioning near scoring opportunities, and defensive movements that restrict dangerous space or passing options. In addition, by adjusting the reward weight for attacking-third ball occupancy, we showed that the model's local action rankings can be inspected under different tactical reward preferences. Because the learned value basis remains fixed during this analysis, the weight adjustment provides a post-hoc means of examining how tactical preferences affect local action rankings.

Overall, the results suggest that the proposed learning formulation as a whole produces more context-sensitive off-ball action rankings than the independent RL baseline in the evaluated data and provides an interpretable basis for analyzing on-ball, off-ball, and defensive decisions. However, because the two approaches differ in several components beyond the MPE-inspired formulation itself, the observed differences should not be interpreted as isolating the effect of the MPE-inspired component.

Several limitations remain. First, although the proposed formulation changed the relationship between team-level average Q-values and season-level xG from negative to weakly positive, the correlation remained weak. The external validity of the learned action values should therefore be examined using additional team-level and possession-level indicators, such as goals, shot creation, possession value, and defensive outcomes. Second, the current state representation mainly relies on player and ball positions, velocities, and derived spatial features. Incorporating body orientation, field of view, and player-specific technical attributes could improve the resolution of tactical evaluation. Third, the reward weights were manually specified in this study. Future work should investigate inverse reinforcement learning to infer reward weights from expert

demonstrations or team-level tactical objectives. Fourth, the current evaluation was conducted on J1 League data. Applying the framework to other leagues, competitions, and data providers would be important for verifying its generalizability. Fifth, the current implementation does not explicitly solve for or verify a mutually consistent MPE strategy profile; MPE is used as a motivating dynamic-game framework, while continuation values are estimated from observed state-action transition data. Finally, the proposed method and the independent RL baseline differ in model architecture, reward representation, learning objective, and training procedure. Consequently, the current experiments do not isolate the contribution of the MPE-inspired component from these other design choices. Controlled ablation studies using a common backbone and training setup are required to quantify the contribution of each component separately.

In future work, we aim to extend the proposed framework toward a more flexible tactical analysis system. In particular, richer state representations, data-driven reward estimation, controlled ablation studies, and validation across multiple competitions will be necessary to establish the robustness of the learned action values. Combining the model with match-video visualization could also support coaches, analysts, and viewers in understanding why a given movement is valuable within the strategic context of a football possession.

Acknowledgements. The J.League dataset used in this study was provided by the Research Organization of Information and Systems, The Institute of Statistical Mathematics and Data Stadium Inc. We are grateful to the following supporters of our crowdfunding project (names given in no particular order): Eiji Konaka, Junya Yatabe, Yu Ukai, Shotaro Ishihara, Narita Youkan, Takashi Hamaji, Atom Scott, Daichi Ishii, Ikuma Uchida, Yasuyuki Sawada, Yosuke Yasuda, Yumu Tabuchi, Hiroyuki Shindo, Kazuyuki Saka, Red Dot Drone Japan, Yuichiro Someya, Keita Inoue, Masato Max Yamamoto, Rikuhei Umemoto, Susumu Fukutome, Takao Toshishige, Makoto Isono, Sho Yokoyama, Leo Nakamura, Masayuki Fukui, and Hayato Ohtani. The author K.I. would like to take this opportunity to thank the "THERS Make New Standards Program for the Next Generation Researchers."

Declarations

- Funding: This work was financially supported by JST SPRING, Grant Number JPMJSP2125. This work was supported by JSPS KAKENHI 23H03282.
- Competing interests: The authors have no competing interests to declare that are relevant to the content of this article.
- Ethics approval: Not applicable
- Consent to participate: Not applicable
- Consent for publication: Not applicable
- Availability of data and materials: The event and tracking data used in this study are subject to the terms and licenses of the original data providers and cannot be redistributed by the authors. The released code provides the implementation and experiment scripts, but users must obtain the underlying data from the corresponding data providers under their own licenses.

- Code availability: The source code for the model implementation, training scripts, experiment scripts, and documentation needed to run the experiments will be made publicly available upon acceptance.
- Authors' contributions: K.I., T.S., K.K., and K.F. contributed to the study conception and design. Data preparation, modeling, and analysis were performed by K.I. The first draft of the manuscript was written by K.I., and all authors commented on previous versions of the manuscript. All authors read and approved the final manuscript.

References

[1] Link, D., Lang, S. & Seidenschwarz, P. Real time quantification of dangerousity in football using spatiotemporal tracking data. *PloS one* **11**, e0168768 (2016).

[2] Taki, T., Hasegawa, J.-i. & Fukumura, T. Development of motion analysis system for quantitative evaluation of teamwork in soccer games. *Proceedings of 3rd IEEE International Conference on Image Processing* **3**, 815–818 (1996).

[3] Decroos, T., Bransen, L., Van Haaren, J. & Davis, J. Actions speak louder than goals: Valuing player actions in soccer 1851–1861 (2019).

[4] Simpson, I., Beal, R. J., Locke, D. & Norman, T. J. Seq2event: Learning the language of soccer using transformer-based match event prediction. *Proceedings of the 28th ACM SIGKDD conference on knowledge discovery and data mining* 3898–3908 (2022).

[5] Teranishi, M., Tsutsui, K., Takeda, K. & Fujii, K. Evaluation of creating scoring opportunities for teammates in soccer via trajectory prediction. *International Workshop on Machine Learning and Data Mining for Sports Analytics* 53–73 (2022).

[6] Nakahara, H., Tsutsui, K., Takeda, K. & Fujii, K. Action valuation of on-and off-ball soccer players based on multi-agent deep reinforcement learning. *IEEE Access* **11**, 131237–131244 (2023).

[7] Ide, K., Someya, T., Kawaguchi, K. & Fujii, K. Interpretable low-dimensional modeling of spatiotemporal agent states for decision making in football tactics. *Proceedings of the 12th International Conference on Sport Sciences Research and Technology Support (icSPORTS)* (2024).

[8] Rahimian, P., Van Haaren, J., Abzhanova, T. & Toka, L. Beyond action valuation: A deep reinforcement learning framework for optimizing player decisions in soccer. *16th Annual MIT Sloan Sports Analytics Conference. Boston, MA, USA: MIT* 1–25 (2022).

[9] Ide, K., Someya, T., Kawaguchi, K. & Fujii, K. Expandable decision-making states for multi-agent deep reinforcement learning in soccer tactical analysis.

Journal of Big Data (2026).

[10] Adusumilli, K. & Eckardt, D. Temporal-difference estimation of dynamic discrete choice models. *The Review of Economic Studies* **93**, 2181–2214 (2026).

[11] Robberechts, P. & Davis, J. How data availability affects the ability to learn good xg models. *International Workshop on Machine Learning and Data Mining for Sports Analytics* 17–27 (2020).

[12] Power, P., Ruiz, H., Wei, X. & Lucey, P. Not all passes are created equal: Objectively measuring the risk and reward of passes in soccer from tracking data. *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining* 1605–1613 (2017).

[13] Fernández, J., Bornn, L. & Cervone, D. Decomposing the immeasurable sport: A deep learning expected possession value framework for soccer. *13th MIT Sloan Sports Analytics Conference* **2** (2019).

[14] Singh, K. Introducing expected threat (xt): Modelling team behaviour in possession to gain a deeper understanding of buildup play. *Karun Singh* (2018). URL https://karun.in/blog/expected-threat.html.

[15] Rudd, S. A framework for tactical analysis and individual offensive production assessment in soccer using markov chains. *New England Symposium on Statistics in Sports* (2011).

[16] Fujimura, A. & Sugihara, K. Geometric analysis and quantitative evaluation of sport teamwork. *Systems and Computers in Japan* **36**, 49–58 (2005).

[17] Brefeld, U., Lasek, J. & Mair, S. Probabilistic movement models and zones of control. *Machine Learning* **108**, 127–147 (2019).

[18] Fernandez, J. & Bornn, L. Wide open spaces: A statistical technique for measuring space creation in professional soccer. *Sloan sports analytics conference* **2018** (2018).

[19] Umemoto, R., Tsutsui, K. & Fujii, K. Location analysis of players in uefa euro 2020 and 2022 using generalized valuation of defense by estimating probabilities. *Proceedings of the 13th International Conference on Sport Sciences Research and Technology Support (icSPORTS)* (2025).

[20] Toda, K., Teranishi, M., Kushiro, K. & Fujii, K. Evaluation of soccer team defense based on prediction models of ball recovery and being attacked: A pilot study. *Plos one* **17**, e0263051 (2022).

[21] Yeung, C., Sit, T. & Fujii, K. Transformer-based neural marked spatio temporal point process model for analyzing football match events. *Applied Intelligence* **55**,

335 (2025).

[22] Le, H. M., Yue, Y., Carr, P. & Lucey, P. Coordinated multi-agent imitation learning. *International Conference on Machine Learning* 1995–2003 (2017).

[23] Fujii, K., Takeishi, N., Kawahara, Y. & Takeda, K. Decentralized policy learning with partial observation and mechanical constraints for multiperson modeling. *Neural Networks* **171**, 40–52 (2024).

[24] Fujii, K. *et al.* Estimating counterfactual treatment outcomes over time in complex multiagent scenarios. *IEEE Transactions on Neural Networks and Learning Systems* **36**, 2103–2117 (2024).

[25] Wang, Z. *et al.* Tacticai: an ai assistant for football tactics. *Nature communications* **15**, 1906 (2024).

[26] Kurach, K. *et al.* Google research football: A novel reinforcement learning environment. *Proceedings of the AAAI Conference on Artificial Intelligence* **34**, 4501–4510 (2020).

[27] Scott, A., Fujii, K. & Onishi, M. How does ai play football? an analysis of rl and real-world football strategies. *Proceedings of the 14th International Conference on Agents and Artificial Intelligence (ICAART)* **1**, 42–52 (2022).

[28] Fujii, K. *et al.* Adaptive action supervision in reinforcement learning from real-world multi-agent demonstrations. *16th International Conference on Agents and Artificial Intelligence (ICAART' 24)* **2**, 27–39 (2024).

[29] Fujii, K. *Machine Learning in Sports: Open Approach for Next Play Analytics* (Springer, 2025).

[30] Lin, F., Huang, S., Pearce, T., Chen, W. & Tu, W.-W. Tizero: Mastering multi-agent football with curriculum learning and self-play. *Proceedings of the 2023 International Conference on Autonomous Agents and Multiagent Systems* 67–76 (2023).

[31] Liu, G. & Schulte, O. Deep reinforcement learning in ice hockey for context-aware player evaluation. *International Joint Conference on Artificial Intelligence* (2018).

[32] Ding, N., Takeda, K. & Fujii, K. Deep reinforcement learning in a racket sport for player evaluation with technical and tactical contexts. *IEEE Access* **10**, 54764–54772 (2022).

[33] Rahimian, P. & Toka, L. Inferring the strategy of offensive and defensive play in soccer with inverse reinforcement learning. *Machine Learning and Data Mining for Sports Analytics: 8th International Workshop, MLSA 2021, Virtual Event,*

September 13, 2021, Revised Selected Papers 26–38 (2022).

[34] Luo, Y. Inverse reinforcement learning for team sports: Valuing actions and players (2020).

[35] Palacios-Huerta, I. Professionals play minimax. *The Review of Economic Studies* **70**, 395–415 (2003).

[36] Chiappori, P.-A., Levitt, S. & Groseclose, T. Testing mixed-strategy equilibria when players are heterogeneous: The case of penalty kicks in soccer. *American Economic Review* **92**, 1138–1151 (2002).

[37] Bar-Eli, M. & Azar, O. H. Penalty kicks in soccer: An empirical analysis of shooting strategies and goalkeepers' preferences. *Soccer & Society* **10**, 183–191 (2009).

[38] Auer, B. R. & Hiller, T. On the evaluation of soccer players: a comparison of a new game-theoretical approach to classic performance measures. *Applied Economics Letters* **22**, 1100–1107 (2015).

[39] Chacoma, A., Almeira, N., Perotti, J. I. & Billoni, O. V. Stochastic model for football's collective dynamics. *Physical Review E* **104**, 024110 (2021).

[40] Ferrarini, A. A new game theory algorithm simulates soccer matches: Reducing complexity to its irreducible essence. *Selforganizology* **1**, 211–218 (2014).

[41] Beal, R., Chalkiadakis, G., Norman, T. & Ramchurn, S. D. Optimising game tactics for football. *Adaptive Agents and Multi-Agent Systems* (2020).

[42] Lennartsson, J., Lidström, N. & Lindberg, C. Game intelligence in team sports. *PloS one* **10**, e0125453 (2015).

[43] Milani Fard, A., Salmani, V., Naghibzadeh, M., Khajouie Nejad, S. & Ahmadi, H. Game theory-based data mining technique for strategy making of a soccer simulation coach agent. *Information systems technology and its applications–6th international conference–ISTA 2007* 54–65 (2007).

Supplementary Materials

A State Representation

This appendix provides the detailed definitions of the state representations used in this study. We compare two types of state representations: Position and Velocity State (PVS) and Expandable Decision-Making State (EDMS). PVS is used as a baseline representation that directly encodes the physical configuration of players and the ball. In contrast, EDMS is designed to incorporate tactically interpretable features, such as spatial value, defensive pressure, passing options, and shooting opportunities.

A.1 PVS: Position and Velocity State

The Position and Velocity State (PVS) is used as the baseline state representation. Following previous reinforcement learning-based approaches for football action valuation, PVS represents the game situation using only the physical information of all players and the ball. Specifically, for each time step t, the state consists of the two-dimensional position and velocity of the ball and all players on the pitch.

Let $p_{i,t} = (x_{i,t}, y_{i,t})$ and $v_{i,t} = (\dot{x}_{i,t}, \dot{y}_{i,t})$ denote the position and velocity of player i at time t, respectively. Similarly, let $p_{b,t} = (x_{b,t}, y_{b,t})$ and $v_{b,t} = (\dot{x}_{b,t}, \dot{y}_{b,t})$ denote the position and velocity of the ball. The PVS representation is then defined as

$$s_t^{\mathrm{PVS}} = [p_{b,t}, v_{b,t}, p_{1,t}, v_{1,t}, \ldots, p_{N,t}, v_{N,t}], \tag{19}$$

where N is the number of players considered in the model.

In this study, the attacking direction is normalized so that the attacking team always moves from left to right. This normalization allows the model to learn action values under a consistent spatial coordinate system across different possessions and matches.

Although PVS provides a compact and direct description of the physical configuration of the game, it does not explicitly encode tactical relationships among players. For example, quantities such as available space, defensive pressure, passing lanes, and shooting angles must be inferred indirectly from raw coordinates and velocities. Therefore, while PVS is useful as a baseline representation, it may be limited in terms of sample efficiency and interpretability when modeling complex tactical decision-making in football.

A.2 EDMS: Expandable Decision-Making State

The Expandable Decision-Making State (EDMS) is the proposed state representation used in this study. EDMS is designed to represent football-specific tactical information in an interpretable and modular form. Unlike PVS, which directly uses raw position and velocity information, EDMS transforms the physical configuration of the pitch into a set of semantically meaningful variables related to space, pressure, passing opportunities, shooting opportunities, and possession context.

The term "expandable" indicates that the state representation is not restricted to a fixed set of variables. Instead, EDMS is organized as a modular structure, allowing

additional tactical or provider-specific variables to be incorporated without changing the overall reinforcement learning framework. This property is important for practical football analytics, where the available data and the tactical concepts of interest may differ depending on the league, data provider, or analysis objective.

EDMS consists of two major components: Absolute State and Relative State. The Absolute State contains global information that is not defined with respect to a specific target player. The Relative State contains player-specific information that changes depending on the evaluated player and the possession context. The Relative State is further divided into On-ball State and Off-ball State, depending on whether the target player is in possession of the ball.

A.2.1 Absolute State

The Absolute State describes global contextual information of the current game situation. In this study, it includes the distance from the ball to each team's offside line and the team formation.

The offside line is defined according to the attacking direction and the positions of the defending players. When all attacking players are located in their own half, the halfway line is used as the offside line. Otherwise, the offside line is defined by the second-deepest defending player relative to the attacking direction. The distance from the ball to this line represents how much vertical space is available behind the defensive line.

The formation variable represents the nominal team formation provided in the data. This variable gives coarse tactical context regarding the team structure. In the current implementation, formation changes caused by substitutions or in-game tactical adjustments are not explicitly modeled.

A.2.2 Off-ball State

The Off-ball State represents the tactical situation of an attacking player who does not possess the ball. For each off-ball player, EDMS computes variables that describe passing accessibility, defensive pressure, and spatial value.

First, the distance between the ball and the off-ball player is computed as

$$d_{i,t}^{\mathrm{ball}} = \|p_{i,t} - p_{b,t}\|_2. \tag{20}$$

This variable captures how easily the player can be involved in the current possession.

Second, EDMS computes the time required for the nearest opposing player to reach the off-ball player. This variable is denoted as

$$\tau_{i,t}^{\mathrm{player}}. \tag{21}$$

A larger value indicates that the player is under weaker defensive pressure and has more time to receive or control the ball.

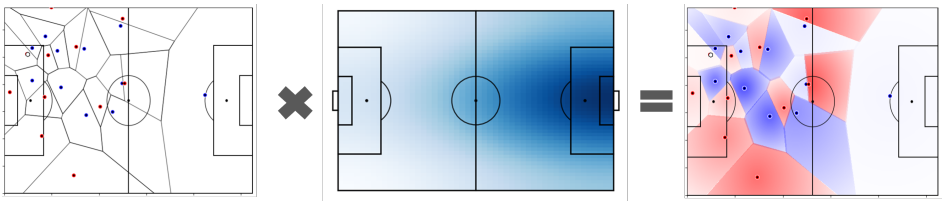


Fig. 9 This figure illustrates the procedure for calculating the space score. From left to right, it shows: the Voronoi diagram considering each player's position and velocity; the importance of each area on the pitch; and the final space score distribution, which is the product of the two. The area importance is modeled based on proximity to the opponent's goal and distance from the center of the pitch, and it is maximized in the central area in front of the opponent's goal

Third, EDMS computes the time required for the nearest opposing player to reach the passing lane from the ball to the off-ball player. This variable is denoted as

$$\tau_{i,t}^{\text{passline}}. \tag{22}$$

A larger value indicates that the passing lane is less likely to be intercepted.

Fourth, EDMS includes a space score, which quantifies the value of the space occupied or controlled by the player. The space score is computed by combining the player's space control with the tactical importance of the corresponding pitch area. The spatial importance is designed to be higher near the opponent's goal and in central areas of the pitch. Thus, the space score assigns high values to regions that are both controllable and tactically dangerous.

Fifth, EDMS includes the change in space score when the player moves one meter in each of eight directions. Let $\mathcal{D}$ denote the set of eight movement directions:

$$\mathcal{D} = \{\text{Right}, \text{Up-Right}, \text{Up}, \text{Up-Left}, \text{Left}, \text{Down-Left}, \text{Down}, \text{Down-Right}\}. \tag{23}$$

For each direction $d \in \mathcal{D}$, the change in space score is defined as

$$\Delta\text{SpaceScore}_{i,t}^{d} = \text{SpaceScore}(p_{i,t} + u_d) - \text{SpaceScore}(p_{i,t}), \tag{24}$$

where u_d is a one-meter displacement vector in direction d.

Finally, EDMS defines a pass score that summarizes the quality of a potential pass to the off-ball player. The pass score is computed as a weighted sum of distance, space value, pressure on the receiver, and pressure on the passing lane:

$$\text{PassScore}_{i,t} = k_1 d_{i,t}^{\text{ball}} + k_2 \text{SpaceScore}_{i,t} + k_3 \tau_{i,t}^{\text{player}} + k_4 \tau_{i,t}^{\text{passline}}. \tag{25}$$

In this study, the weights are set as

$$k_1 = 0.5, \quad k_2 = 0.3, \quad k_3 = 0.2, \quad k_4 = 0.2. \tag{26}$$

These weights were determined based on preliminary analysis of pass success prediction.

A.2.3 On-ball State

The On-ball State represents the tactical situation of the player in possession of the ball. It is divided into Intra-Possession State and Inter-Possession State. The Intra-Possession State describes situations in which the same team maintains possession, while the Inter-Possession State describes situations involving a possible transition of possession between teams.

The Intra-Possession State includes variables related to pressure, goal proximity, dribbling opportunities, shooting opportunities, and long-ball options.

First, EDMS computes the time required for the nearest opposing player to reach the ball:

$$\tau_t^{\text{ball}}. \tag{27}$$

This variable represents the pressure applied to the ball carrier.

Second, the distance from the ball carrier to the opponent's goal is computed as

$$d_t^{\text{goal}} = \|p_{b,t} - p^{\text{goal}}\|_2, \tag{28}$$

where p^{goal} denotes the center of the opponent's goal.

Third, the angle from the ball carrier to the opponent's goal is computed. This angle captures the geometric quality of the shooting or attacking situation.

Fourth, EDMS defines a dribble score. The dribble score is computed as the change in space score when the ball carrier moves one meter in each of the eight directions:

$$\text{DribbleScore}_t^d = \text{SpaceScore}(p_{b,t} + u_d) - \text{SpaceScore}(p_{b,t}), \quad d \in \mathcal{D}. \tag{29}$$

This variable evaluates whether a short movement by the ball carrier would improve the tactical situation.

Fifth, EDMS includes a shot score. The shot score evaluates the quality of a potential shot by considering the angle to the goal and the probability that defenders block the shooting lane. Only defenders located inside the triangular region formed by the ball carrier and the two goalposts are considered, excluding the goalkeeper. The blocking probability is aggregated over the goal mouth, and the shot score is computed only when the ball carrier is within 30 meters of the opponent's goal.

Sixth, EDMS includes a long-ball score. The pitch is divided into left, central, and right zones, and the zone containing the tallest attacking player is assigned a high value. This variable allows the model to represent long-ball options toward aerially dominant players.

A.2.4 Inter-Possession State

The Inter-Possession State represents situations in which possession may change between teams. This component is used to model transitional situations, such as turnovers, loose balls, and defensive recoveries.

The Inter-Possession State includes the following variables. First, EDMS computes the time required for the nearest player from each team to reach the ball. This variable captures which team is more likely to gain possession.

Table 5 Summary of the EDMS components.

Category	Subcategory	Main variables
Absolute State	–	Offside-line distance, formation
Relative State	Off-ball State	Ball distance, pressure, pass-line pressure, space score, pass score
Relative State	On-ball / Intra-Possession State	Ball pressure, goal distance, goal angle, dribble score, shot score, long-ball score
Relative State	On-ball / Inter-Possession State	Time to reach ball, goal distance, goal angle, ball velocity, transition indicator

Second, EDMS computes the distance from the ball to each team's goal:

$$d_{c,t}^{\text{goal}} = \|p_{b,t} - p_c^{\text{goal}}\|_2, \quad c \in \{\text{attack}, \text{defense}\}. \tag{30}$$

Third, EDMS computes the angle from the ball to each team's goal. Fourth, the velocity of the ball is included to capture whether the ball is moving toward a dangerous area or away from pressure. Fifth, a transition indicator is included. This variable is set to 1 when the player is switching from attack to defense and 0 when the player is switching from defense to attack.

A.2.5 Summary of EDMS

Table 5 summarizes the main components of EDMS.

Compared with PVS, EDMS explicitly encodes tactical concepts that are relevant to football decision-making. This allows the reinforcement learning model to evaluate not only the physical location of players but also the meaning of their positions in the context of pressure, space, passing options, and goal threat. Moreover, because EDMS is modular, additional variables such as body orientation, player-specific passing ability, score difference, or set-piece-specific features can be incorporated in future extensions.