Multimodal Pass Receiver Prediction in Soccer with Large Language Models

Tim Schlippe
IU International University of Applied Sciences
Germany
tim.schlippe@iu.org

Frederic Pieper
IU International University of Applied Sciences
Germany
frederic.pieper@iu-study.org

***Abstract*—This paper investigates the potential of Large Language Models (LLMs) for predicting pass receivers in soccer. Through systematic experimentation, we explore which combinations of information types and representations improve prediction accuracy. We developed a four-stage experiment comparing different modalities (*image*, *text*, *their combination*) and additional information (*match context*, *player statistics*, *sequence descriptions*) using five state-of-the-art LLMs. Our results demonstrate that LLMs possess basic tactical understanding applicable to pass receiver prediction. The highest average accuracy (33%) was achieved by combining the *annotated image*, *image description*, *sequence before the pass*, *and player information*—significantly higher than random chance (23%). At the individual model level, Claude 3.5 Sonnet, DeepSeek-V3, GPT-4o, and Llama 3.2 each reached 40% accuracy with their optimal input combinations. Interestingly, additional information does not always improve results; incorporating general match context reduced accuracy from 33% to 28%. LLMs still trail behind human performance (55% accuracy). But our work delivers concrete insights into multimodal input optimization for LLM-based sports predictions, establishing both a public benchmark dataset and a foundation for future research in AI-assisted soccer analytics.**

***Index Terms*—Large Language Models, Multimodal Learning, Pass Receiver Prediction, Sports Analytics, Soccer**

I. Introduction

This section presents the motivation behind our research on pass receiver prediction, outlines our research questions, and describes the structure of this paper.

A. Motivation

In recent years, soccer has undergone a profound transformation driven by comprehensive data collection and analysis. Event and tracking data, supplemented by additional data types, offer new perspectives on game dynamics, player performance, and tactical patterns [1]. One of the central challenges in modern soccer analysis is the examination and prediction of player decisions. In the dynamic context of a soccer match, players often have fractions of seconds to make complex decisions—whether to pass, dribble, or shoot [2].

Passes play a fundamental role in soccer strategy as the most frequent and dynamic component of gameplay. The ability to effectively anticipate passes is often a decisive success factor in high-level competitions [3]. A successful methodology for pass receiver prediction opens up far-reaching application possibilities: coaches could determine player and team predictability, viewer experience could be improved through intelligent camera guidance [4], and referees could be better positioned for possible game situations.

Previous scientific work on this topic shows an exciting development: starting with simple, rule-based approaches [5], progressing to classical machine learning methods [4], and advancing to deep learning architectures [6], particularly transformer models [7]. Despite these advances, pass receiver prediction remains challenging due to the high complexity of the game and the multitude of possible decision options [2].

In parallel, the world of AI has undergone an equally dynamic transformation with the emergence of Large Language Models (LLMs). LLMs are characterized by their ability to respond to new tasks without specific training [8]. Another milestone is the transition to Multimodal Large Language Models (MLLMs), which can process not only text but also other modalities such as images or audio [9]. This opens up entirely new application possibilities, especially in areas where visual information plays a central role—like soccer.

This paper addresses precisely this complex prediction task: It examines the extent to which modern LLMs and particularly multimodal models are capable of predicting the next pass receiver in soccer using zero-shot learning without specialized fine-tuning. For coaching staffs, accurate pass prediction tools could transform match preparation and in-game decision-making by revealing player tendencies, supporting opposition analysis, and enabling more targeted tactical adjustments both during training and competitive matches.

B. Research Questions

1) What is better for pass receiver prediction—*image*, *textual image description*, or their *combination*?
2) Can *additional text information* improve pass receiver prediction?
3) How do different LLMs differ in pass receiver prediction?
4) How good is the pass receiver prediction of LLMs compared to *humans*?

C. Structure of this paper

The remainder of this paper is structured as follows: Section II presents related work on pass receiver prediction and LLMs. Section III describes our experimental setup, including the

specific LLMs evaluated and our custom-built corpus. Section IV details our experimental results across four progressive experiments. Finally, Section V summarizes our findings and outlines promising directions for future research.

II. Related Work

This section reviews existing approaches to pass receiver prediction, available data sources, and the capabilities of LLMs in soccer applications.

A. Methods for Pass Receiver Prediction

Current approaches to pass receiver prediction in soccer show significant methodological diversity, broadly categorizable into classical machine learning models, deep learning approaches, and other methods.

Traditional machine learning models like Random Forest [10], [11], Gradient Boosting Decision Trees [4], and Binary Logistic Regression [12] appear in several studies. Deep learning approaches dominate the literature, including neural networks [2], [6], [10], [13], [14] as well as graph-based approaches such as Graph Neural Networks [15] and Temporal Graph Networks [16]. Other approaches include rule-based systems [5], Inductive Logic Programming [17], and physics-based models [18].

The accuracy of these methods ranges from 29% [17] to 86% [15], with average accuracy of 57%. Many approaches show strong dependencies on training sets and fine-tuning, while LLMs can be applied through simple zero-shot learning.

B. Large Language Models for Soccer Applications

The application of LLMs in soccer is an emerging research field explored to a limited extent so far. Unlu [19] demonstrated that training small models with soccer-related data enables the development of specialized language models. Schilling et al. [20] developed a Retrieval-Augmented Generation (RAG) approach to provide contextual answers to soccer-related questions.

Caron & Müller [21] developed TacticalGPT, an AI-based assistant coach that can understand and predict tactical decisions in professional soccer using natural language. Pavlovich et al. [22] trained a Llama model to generate automated game commentary. Held et al. [23] created X-VARS, a specialized multimodal LLM for analyzing soccer referee decisions that achieves performance at the level of human referees.

These applications show that LLMs possess a basic understanding of the game and can potentially be used for further applications, though—to the best of our knowledge—no prior work has examined their capabilities for pass receiver prediction.

III. Experimental Setup

This section describes the LLMs, dataset, and evaluation metrics used in our investigation of multimodal pass receiver prediction.

A. Large Language Models

In this study, we evaluated five state-of-the-art LLMs with multimodal capabilities:

1) **Gemini Flash 2.0** (Google): Released December 2024, with 1M token input capacity and support for text, image, video, and audio [24].
2) **DeepSeek-V3** (DeepSeek): Released December 2024, an open-source model with 64k token capacity that can extract text from images [25].
3) **Llama 3.2 11B** (Meta): Released September 2024, with 128k token capacity and ability to process text and images [26].
4) **Claude 3.5 Sonnet** (Anthropic): Released June 2024, with 200k token capacity and strong visual processing capabilities [27].
5) **GPT-4o** (OpenAI): Released May 2024, a multimodal model with 128k token capacity that can process text, images, audio, and video [28].

These models were selected based on their state-of-the-art performance, accessibility through chat interfaces, and multimodal capabilities.

B. Our Pass Receiver Prediction Corpus

Our *Pass Receiver Prediction Corpus* is based on a selection of 20 pass scenes from games in the German Bundesliga. These scenes were identified and extracted by manually viewing publicly available video material, particularly highlight compilations. The aim was to create a diverse collection of game scenes that vary both in terms of the game situation and structural properties (e.g., number of passing options, which ranged from 2-7).

The dataset statistics show a balanced distribution of game situations: 45% of scenes occur in the first half and 55% in the second half, with 40% taking place in the team's own half and 60% in the opponent's half. The average number of visible passing options per scene is 4.8, with a standard deviation of 1.5. For each scene, the ground truth pass receiver was manually annotated by reviewing the complete video sequence.

The corpus consists of several components:

1) **Visual data**: Images of the game scenes with varying levels of annotation (from raw images to fully annotated with player names, numbers, and bounding boxes)
2) **Textual scene descriptions**: Structured descriptions of each scene covering player positioning, teammate network, defensive configuration, and spatial dynamics.
3) **Player statistics**: Data on the ball-carrying players' passing abilities and preferences
4) **Match context**: Information about game time, score, team positions in the league standings
5) **Sequence descriptions**: Textual descriptions of the movement sequences immediately before the image was taken

This systematic combination of visual and textual data enables a comprehensive and multidimensional analysis of soccer game sequences. Figure 1 illustrates a typical textual

The ball carrier (Brandt, #10) is positioned just outside the central penalty area, approximately 20 meters from the goal, slightly left of center. He's moving forward with controlled speed parallel to the penalty area edge, maintaining close ball control. One opponent is approaching from the front at a moderate pace, about 3 meters away. Potential receivers include: Guirassy (#9) positioned centrally in the penalty area 7 meters from goal, moving slightly left to create separation from defenders; Gittens (#43) positioned to Brandt's right about 10 meters away, moving forward into the penalty area; and Beier (#14) stationed centrally about 15 meters from goal without significant movement.

Fig. 1: Example of a textual scene description.

Statistic for Ball Carrier (Julian Brandt)	Per 90	Percentile
Progressive Passes	5.64	88
Through Balls	0.54	90
Passes into Penalty Area	1.59	60
...	...	...

Fig. 2: Example of player statistics (selected key metrics).

scene description from our corpus, detailing player positions, movements, and spatial relationships. Figure 2 shows selected player statistics for the ball carrier, where "Per 90" represents the frequency per complete match and "Percentile" indicates how the player ranks relative to others in the same position (90th percentile means better than 90% of players). The complete dataset is publicly available on GitHub[1].

C. Evaluation Metrics

We used accuracy as the metric for evaluating model performance, defined as the proportion of correctly predicted pass receivers out of the total number of analyzed game scenes.

Accuracy was calculated for each LLM individually, and model-wide average accuracy was also determined.

IV. Experiments and Results

This section presents our experimental methodology and findings on using LLMs for pass receiver prediction across different input modalities and information combinations.

A. Overview of the Experiments

We conducted a series of four progressive experiments:

1) **Optimizing Visual Representations with Annotation Granularity**: Finding which visual annotation of the image is best suited for prediction.
2) **Comparing Modalities: Image, Description, and Multimodal Fusion**: Finding the optimal representation form among image, image description, and their combination.
3) **Evaluating the Impact of Textual Additions**: Determining the added value of individual textual additional information.
4) **Synergizing Modalities: Optimal Text Composition and Visual Representation**: Finding the optimal composition of textual additional information in combination with the optimal visual representation form.

Table I provides a comprehensive overview of the results across all experiments. The table shows accuracy percentages for each model under different input conditions, highlighting the best performance for each model and experiment. As shown, we achieved a maximum average accuracy of 33% when combining *image*, *image description*, *sequence before the pass*, and *player information*. This represents a significant improvement over the random baseline (23%), though it remains below human performance (55%). In the following subsections, we analyze each experiment in detail, exploring the factors that influence model performance and the optimal conditions for pass receiver prediction.

[1]https://github.com/FredDec97/Pass-Receiver-Prediction-with-LLMs

B. Optimizing Visual Representations with Annotation Granularity

In this experiment, we tested seven different annotation levels, from raw images to fully annotated images with player names, numbers, and bounding boxes for both teams. The results showed that the annotation level "*Boxes + Name + Numbers (Ball Team)*" (see Figure 3) with an average accuracy of 22% performed best for pass receiver prediction, although this value is slightly below the random accuracy of 23%.

However, individual model performance varied significantly: Claude 3.5 Sonnet achieved the highest accuracy of 40% with the annotation level "*Boxes + Name + Numbers (Ball Team)*", followed by GPT-4o with 35% using "*Boxes (Ball Team)*", DeepSeek-V3 with 30%, while Llama-3.2-11B (20%) and Gemini Flash 2.0 (20%) showed more modest results.

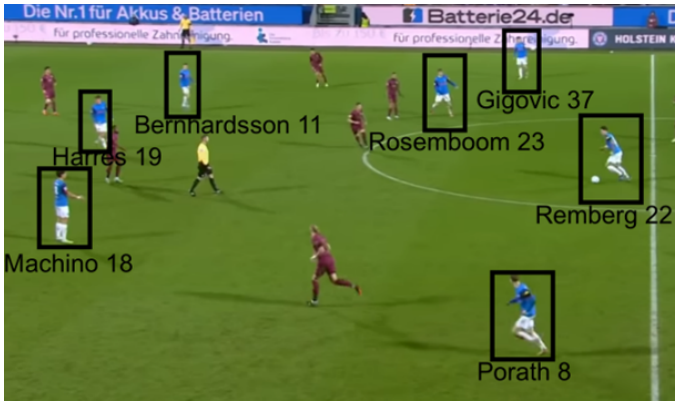


Fig. 3: Optimal Visual Representation.

Some LLMs struggled to autonomously identify the ball-carrying player. Therefore, in subsequent experiments, we explicitly provided this information.

TABLE I: Summary of Pass Receiver Prediction Accuracy (%) across all Experiments

Experiment / Input Type	Claude 3.5	DeepSeek-V3	GPT-4o	Llama 3.2	Gemini 2.0	Average
Experiment 1: Visual Annotations						
Raw Image	15	10	25	5	5	12
Boxes (Ball Team)	25	25	35	5	0	18
Boxes + Numbers (Ball Team)	20	25	25	20	15	21
Boxes + Name + Numbers (Ball Team)	**40**	30	15	20	5	**22**
Boxes (Both Teams)	5	20	25	5	5	12
Boxes + Numbers (Both Teams)	25	10	10	10	5	12
Boxes + Name + Numbers (Both Teams)	15	15	30	0	20	16
Experiment 2: Modality Comparison						
Visual (Optimal Annotation)	**40**	30	15	20	5	22
Textual (Image Description)	25	35	25	20	35	28
Multimodal (Visual + Textual)	35	35	25	20	35	**30**
Experiment 3: Additional Information						
Multimodal Only	35	35	25	20	35	30
+ Game Context	20	35	20	25	30	26
+ Player Information	20	35	25	**40**	30	30
+ Sequence Before Pass	25	**40**	35	30	30	**32**
Experiment 4: Optimal Combination						
Multimodal Only	35	35	25	20	**35**	30
+ Sequence Before Pass	25	**40**	35	30	30	32
+ Sequence + Player Info	30	35	**40**	**40**	20	**33**
+ Sequence + Player + Context	15	35	35	25	30	28
Best Performance per Model	**40**	**40**	**40**	**40**	**35**	**33**
Human Performance (Average)						**55**
Human Performance (Range)						**35-85**

C. Comparing Modalities: Image, Description, and Multimodal Fusion

To answer our first research question—defined in Section I-B, we compared three representation forms:

1) Optimal visual representation (image with annotation level "*Boxes + Name + Numbers (Ball Team)*")
2) Textual scene description (image description)
3) Their combination

Using only textual image description led to a higher average prediction accuracy of 28% compared to 22% for visual information alone. The combination of both modalities achieved the best results with an accuracy of 30%.

This increase shows that multimodal information transfer enables the LLMs to better capture the game scene and make more accurate pass receiver predictions, consistent with findings from Wang et al. [29] that the combination of image and image description performs best in spatial reasoning tasks.

At the individual model level, we observed notable differences in how LLMs responded to different modalities: Claude 3.5 Sonnet performed best with visual information alone (40%), actually declining to 25% with textual descriptions, before recovering to 35% with the multimodal combination. In contrast, Gemini Flash 2.0 showed dramatic improvement when moving from visual-only (5%) to text (30%) or multimodal inputs (35%). DeepSeek-V3 maintained consistent performance (30-35%) across all modalities, while GPT-4o improved from 25% with visual input to 35% with the multimodal combination. Llama 3.2 showed a steady 20% accuracy regardless of the input modality.

D. Evaluating the Impact of Textual Additions

To answer our second research question—defined in Section I-B, we tested the influence of three additional information types on prediction accuracy:

1) Game context (match situation, score, standings)
2) Player information (passing statistics of the ball-carrying player)
3) Sequence before the pass (description of movement before the image was taken)

The sequence before the pass proved most valuable, increasing average accuracy to 32%. *Player information* maintained the baseline accuracy of 30%, while *game context* reduced accuracy to 26%. This suggests that not all contextual information is beneficial, and dynamic information about immediately preceding events contributes most to prediction quality.

The models showed different sensitivities to additional information: Claude 3.5 Sonnet performed worse with any additional information, while Llama 3.2 achieved its best result (40%) with *player information*. DeepSeek-V3 and GPT-4o benefited from the *sequence before the pass*, while Gemini Flash 2.0 maintained consistent performance across conditions.

E. Synergizing Modalities: Optimal Text Composition and Visual Representation

While in the previous section, we compared the impact of the individual text additions, in the final experiment we gradually enriched the average optimal representation form with textual additional information in order of their effectiveness. Adding the *sequence before the pass* increased average accuracy to 32%. Further adding *player information* improved accuracy to 33%—the highest value in the entire experiment series. However, further adding *game context* reduced accuracy to 28%.

The models showed varying responses to information composition: Claude 3.5 Sonnet performed best with minimal addi-

tional information (35%), while GPT-4o and Llama 3.2 benefited significantly from enriched inputs, particularly *player information* (40% each). DeepSeek-V3 showed consistent performance across all variants (35-40%).

F. Human Performance and LLM Top Performers

To benchmark the LLM results and answer the third and fourth research question—defined in Section I-B, we conducted a survey where participants were shown the same 20 scenes of our *Pass Receiver Prediction Corpus* with the optimal information composition determined from our experiments. Our study included 110 participants (66.4% male, 33.6% female), primarily between 24-34 years old (59.1%), followed by 35-44 years (16.4%) and 22-24 years (10.9%). The participants demonstrated strong soccer engagement: 52.7% expressed high interest in soccer, 79% watched soccer games at least once a month, and 71.8% had played soccer for at least one year, with 44.5% having over ten years of playing experience.

Humans achieved an average accuracy of 55%, with individual performances ranging from 35% to 85%. This represents a substantial gap compared to the best overall LLM performance of 33% (achieved with the combination of *image*, *image description*, *sequence before the pass*, and *player information*). At individual model level, highest performances were achieved by Claude 3.5 Sonnet (40% with visual information only), GPT-4o (40% with *image*, *image description*, and *player information*), Llama 3.2 (40% with *image*, *description*, and *player information*), and DeepSeek-V3 (40% with *image*, *image description*, and *sequence before the pass*). These peak performances remain significantly below human average and far from the top human performers who reached 85% accuracy.

For 76%, the images were most helpful, while 22% preferred the combination of image and text. Only one person found the texts alone most helpful. This suggests that visual information plays a particularly important role in human pass prediction, whereas LLMs benefit more from the combination of modalities.

V. Conclusion and Future Work

This section summarizes our key findings on LLM-based pass receiver prediction, discusses their implications for sports analytics, and outlines promising directions for extending this research.

A. Conclusion

Our paper investigated the potential of LLMs for predicting pass receivers in soccer. Through a systematic four-stage experimental approach, we evaluated five state-of-the-art LLMs across different input modalities and information combinations. We first optimized visual representations by testing various annotation levels, then compared image-only, text-only, and multimodal inputs. We further explored the impact of additional contextual information and finally identified optimal combinations of modalities and textual additions. Our key findings include:

1) **LLMs show some capability to interpret soccer scenarios**, enabling them to predict pass receivers with accuracy significantly above random chance (33% vs. 23%). This performance, while modest, suggests that general-purpose LLMs can extract relevant patterns from soccer scenes. The highest accuracy was achieved through a multimodal approach combining *annotated images*, *textual descriptions*, *sequence information before the pass*, and *player information*.
2) **Visual information alone proved insufficient** for optimal predictions, while textual descriptions outperformed visual-only inputs. However, the optimal approach was multimodal, combining visual and textual inputs.
3) **Adding relevant dynamic context** through *descriptions of the sequence before the pass* further improved performance, while *player information* contributed to reaching the peak accuracy. Interestingly, adding *general match context information* reduced accuracy, suggesting not all additional information is beneficial.
4) **Different models showed varied strengths** and sensitivities to input compositions, indicating no universal information composition works optimally for all models.
5) **LLMs still trail human performance substantially** (33% vs. 55% accuracy) despite these improvements.

Our findings highlight both the potential and current limitations of using LLMs for pass receiver prediction.

B. Future Work

While our research demonstrates that zero-shot LLMs can effectively contribute to sports prediction tasks, the performance gap between LLMs (33%) and humans (55%) suggests significant room for improvement. To address the performance gap and to leverage the observed model-specific strengths, we propose the following research directions:

1) **Expanding the dataset** to include more game scenes from different leagues and competitions. A larger and more diverse corpus would improve the statistical robustness and generalizability of results.
2) **Integrating tracking and event data** to provide more precise movement information. These quantitative measurements could significantly enhance the models' understanding of player dynamics and spatial relationships.
3) **Implementing RAG systems** with access to sports-specific knowledge bases. By retrieving relevant tactical patterns, player statistics, and historical match data, LLMs could make more informed predictions based on soccer-specific knowledge beyond their general training.
4) **Developing specialized fine-tuned models** trained specifically on soccer data. While our zero-shot approach demonstrated the inherent capabilities of general-purpose LLMs, domain-specific training could significantly enhance performance for this specialized task.
5) **Using model ensembles** to combine strengths of different LLMs. As results showed, different models excel under different inputs, suggesting a weighted ensemble approach could yield better overall performance.

6) **Leveraging video-capable LLMs** for dynamic scene analysis. Models like Gemini 1.5, GPT-4o, and Claude 3 Opus can process video inputs, allowing analysis of complete temporal sequences rather than static images. This temporal dimension is crucial for understanding player momentum, defensive shifts, and passing opportunities as they develop over time.

These technical advancements aim not only to improve prediction accuracy but also to enable practical applications in professional soccer. Enhanced pass receiver prediction systems could serve as valuable coaching tools for analyzing player decision-making patterns and opponent tendencies. They could also transform broadcast experiences through augmented reality overlays highlighting likely passing options in real-time, or enable more sophisticated performance metrics that quantify the predictability of players and teams. For referees, such systems could support optimal positioning by anticipating likely play development. Collectively, these approaches could further enhance the performance of LLM-supported soccer analyses and advance the integration of AI into sports analytics, potentially narrowing the current gap between model and human performance.

ACKNOWLEDGMENT

This research was supported by the IU International University of Applied Sciences (*IU Incubator*) under the internal funding framework for the period from October 2023 to September 2025.

REFERENCES

[1] J. Davis, L. Bransen, L. Devos, A. Jaspers, W. Meert, P. Robberechts, J. Van Haaren, and M. Van Roy, "Methodology and Evaluation in Sports Analytics: Challenges, Approaches, and Lessons Learned," *Machine Learning*, vol. 113, no. 9, pp. 6977–7010, 2024.

[2] Y. Honda, R. Kawakami, R. Yoshihashi, K. Kato, and T. Naemura, "Pass Receiver Prediction in Soccer Using Video and Players' Trajectories," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2022, pp. 3502–3511.

[3] S. Sanyal, "Who Will Receive the Ball? Predicting Pass Recipient in Soccer Videos," *Journal of Visual Communication and Image Representation*, vol. 78, p. 103190, 2021.

[4] H. Li and Z. Zhang, "Predicting the Receivers of Football Passes," in *Machine Learning and Data Mining for Sports Analytics*, U. Brefeld, J. Davis, J. Van Haaren, and A. Zimmermann, Eds., vol. 11330. Springer International Publishing, 2019, pp. 167–177.

[5] P. Fournier-Viger, T. Liu, and J. Chun-Wei Lin, "Football Pass Prediction Using Player Locations," in *Machine Learning and Data Mining for Sports Analytics*, U. Brefeld, J. Davis, J. Van Haaren, and A. Zimmermann, Eds., vol. 11330. Springer International Publishing, 2019, pp. 152–158.

[6] T. Kaneko, R. Kawakami, T. Naemura, and N. Inoue, "Augmenting Pass Prediction via Imitation Learning in Soccer Simulations," in *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*. IEEE, 2024, pp. 3194–3203.

[7] Y. Pei, V. De Silva, and M. Caine, "Passing Heatmap Prediction Based on Transformer Model and Tracking Data," *arXiv preprint arXiv:2309.01526*, 2023.

[8] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell *et al.*, "Language Models are Few-Shot Learners," in *Advances in Neural Information Processing Systems*, vol. 33, 2020, pp. 1877–1901.

[9] S. Yin, C. Fu, S. Zhao, K. Li, X. Sun, T. Xu, and E. Chen, "A Survey on Multimodal Large Language Models," *National Science Review*, vol. 11, no. 12, p. nwae403, 2024.

[10] P. Felsen, P. Agrawal, and J. Malik, "What Will Happen Next? Forecasting Player Moves in Sports Videos," in *2017 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2017, pp. 3362–3371.

[11] Y. Dauxais and C. Gautrais, "Predicting Pass Receiver in Football Using Distance Based Features," in *Machine Learning and Data Mining for Sports Analytics*, U. Brefeld, J. Davis, J. Van Haaren, and A. Zimmermann, Eds., vol. 11330. Springer International Publishing, 2019, pp. 145–151.

[12] S. Steiner, S. Rauh, M. Rumo, K. Sonderegger, and R. Seiler, "Using Position Data to Estimate Effects of Perceptual Features of Play on Passing Decisions in Soccer," *Current Issues in Sport Science*, vol. 3, p. 009, 2018.

[13] O. Hubáček, G. Šourek, and F. Železný, "Deep Learning from Spatial Relations for Soccer Pass Prediction," in *Machine Learning and Data Mining for Sports Analytics*, U. Brefeld, J. Davis, J. Van Haaren, and A. Zimmermann, Eds., vol. 11330. Springer International Publishing, 2019, pp. 159–166.

[14] R. Goka, Y. Moroto, K. Maeda, T. Ogawa, and M. Haseyama, "Multimodal Shot Prediction Based on Spatial-Temporal Interaction between Players in Soccer Videos," *Applied Sciences*, vol. 14, no. 11, p. 4847, 2024.

[15] B. Paneru, B. Paneru, R. Poudyal, and K. Poudyal, "Enhancing Soccer Pass Receiver Prediction in Broadcast Images Through Advanced Deep Learning Techniques: A Comprehensive Study on Model Optimization and Performance Evaluation," *Journal of Soft Computing Exploration*, vol. 5, no. 2, pp. 115–121, 2024.

[16] P. Rahimian, H. Kim, M. Schmid, and L. Toka, "Pass Receiver and Outcome Prediction in Soccer Using Temporal Graph Networks," in *Machine Learning and Data Mining for Sports Analytics*, U. Brefeld, J. Davis, J. Van Haaren, and A. Zimmermann, Eds., vol. 2035. Springer Nature Switzerland, 2024, pp. 52–63.

[17] V. Vercruyssen, L. D. Raedt, and J. Davis, "Qualitative Spatial Reasoning for Soccer Pass Prediction," in *MLSA@PKDD/ECML*, 2016.

[18] W. Spearman, A. Basye, G. Dick, R. Hotovy, and P. Pop, "Physics-based Modeling of Pass Probabilities in Soccer," in *MIT Sloan Sports Analytics Conference*, 2017.

[19] E. Unlu, "FootGPT: A Large Language Model Development Experiment on a Minimal Setting," *arXiv preprint arXiv:2308.08610*, 2023.

[20] A. Schilling, J. Anurathan, J. Mühlberger, F. Gerschner, M. Rössle, A. Theissler, and M. Klaiber, "Querying Football Matches for Event Data: Towards Using Large Language Models," in *Sports Analytics*, J. S. Dong, M. Izadi, and Z. Hou, Eds., vol. 14794. Springer Nature Switzerland, 2024, pp. 216–227.

[21] M. Caron and O. Müller, "TacticalGPT: Uncovering the Potential of LLMs for Predicting Tactical Decisions in Professional Football," StatsBomb, Tech. Rep., 2023.

[22] R. V. Pavlovich, E. A. Tsybulko, K. N. Zhigunov, A. V. Khelvas, A. A. Gilya-Zetinov, and I. V. Tykhonov, "Soccer Artificial Intelligence Commentary Service on the Base of Video Analytic and Large Language Models," in *2023 31st Telecommunications Forum (TELFOR)*. IEEE, 2023, pp. 1–4.

[23] J. Held, H. Itani, A. Cioppa, S. Giancola, B. Ghanem, and M. Van Droogenbroeck, "X-VARS: Introducing Explainability in Football Refereeing with Multi-Modal Large Language Model," *arXiv preprint arXiv:2404.06332*, 2024.

[24] Google. (2024, 12) Wir Stellen Gemini 2.0 vor: Unser Neues Modell für die Ära der KI-Agenten. [Online]. Available: https://blog.google/intl/de-de/unternehmen/technologie/gemini-2-0/

[25] DeepSeek-AI, A. Liu, B. Feng, B. Xue, B. Wang, B. Wu, C. Lu, C. Zhao, C. Deng, C. Zhang *et al.*, "DeepSeek-V3 Technical Report," *arXiv preprint arXiv:2412.19437*, 2024.

[26] Meta AI. (2024) Llama 3.2: Revolutionizing Edge AI and Vision with Open, Customizable Models. [Online]. Available: https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/

[27] Anthropic. (2024) Claude 3.5 Sonnet. [Online]. Available: https://www.anthropic.com/news/claude-3-5-sonnet

[28] OpenAI. (2024, 5) Hello GPT-4o. [Online]. Available: https://openai.com/index/hello-gpt-4o/

[29] J. Wang, Y. Ming, Z. Shi, V. Vineet, X. Wang, Y. Li, and N. Joshi, "Is A Picture Worth A Thousand Words? Delving Into Spatial Reasoning for Vision Language Models," *arXiv preprint arXiv:2406.14852*, 2024.