Vol. 14 (2026): Proceedings of the 2nd International Conference on Artificial Intelligence, Modern Engineering and Environmental Sustainability (IC-AIMEES 2026)
E-ISSN: 3105-0522 P-ISSN：3105-0514
DOI: https://doi.org/10.70267/ ic-aimees.20260428434

Research and Analysis of NLP Based Motion Understanding Large Language Model from Perception to Cognition

Chenyang Xue*

School of Mathematical Sciences, East China Normal University, Shanghai 200241, China

**Corresponding author: Chenyang Xue.*

Abstract

With the rapid development of Large Language Models (LLMs), Natural Language Processing (NLP) has paved a new path for intelligent analysis in the field of sports. However, the comprehension of sports knowledge in most existing models is limited to static texts and struggles to capture real-time information. Based on the main line of "from perception to cognition", this paper systematically reviews the research progress of large language models in sports intelligence analysis. At the perception level, this paper reviews the text-based encoding of sports knowledge, video-based understanding of sports events and sports perception-oriented datasets. At the cognition level, this paper investigates the current state of large language models in sports modeling, focusing on tactical analysis, decision understanding, and evaluation, and match trend prediction. Furthermore, this review summarizes the current challenges in data resources and model capabilities, and looks forward to the future development pathways, including multimodal datasets construction, temporal awareness enhancement, and reasoning stability improvement. Therefore, promoting large language models from perception to cognition is expected to realize their in-depth application in the field of sports, becoming an intelligent tool for match comprehension and decision support.

Keywords

large language models, perception to cognition, multi-modal, intelligence analysis

1. Introduction

Natural Language Processing is committed to the effective interaction between computers and natural languages and its evolution has undergone a phase transition from rule-based to statistical-based, and then to deep learning [1]. In 2017, the proposal of the Transformer architecture marked a milestone breakthrough, with its attention mechanism enabling the model to capture long-distance semantic dependencies [2]. Based on this, BERT uses an encoder architecture to process the corpus in batches by the method of pre-training and fine-tuning; GPT, in contrast, adopts a decoder architecture, demonstrating powerful text generation capability [1]. In recent years, the emergence of various large language models such as ChatGPT, DeepSeek, and Gemini has further promoted the development of NLP, providing important support for machine translation, text creation, sentiment analysis, and other aspects.

Meanwhile, the demand for intelligent analysis in modern sports is becoming increasingly urgent. However, the essence of sports is a non-verbal multi-modal interaction scene, which has the characteristics of high dynamic, strong temporal dependencies, and multimodality. Taking a football match as an example, players' movement, tactical analysis, and instantaneous decision-making are difficult to record in text form. Traditional sports data analysis mostly relies on data statistics or manual video annotation, failing to deal with real-time match information, which constitutes the core challenge of the application of large language models in the sports field. To overcome these limitations, researchers began to combine large language models with multi-modal large models equipped with visual comprehension capabilities, attempting to establish a pathway from perception to cognition, so that the model can not only visually understand the game, but also comprehend the physical laws and adversarial strategies behind the game.

Specifically, in cognitive science, perception and cognition constitute the two basic levels of human contact and understanding of the world. Perception is the process of receiving external stimuli through the senses and forming preliminary representations, and cognition is the understanding of the world formed by information processing and reasoning based on perception. Mapped to the sports scene, the memory of sports knowledge and the capture of visual information such as player's movements and running tracks belong to the level of perception; on this basis, the processing of the acquired information, such as understanding the tactical intention, evaluating the quality of decision-making and predicting the trend of the game, belong to the level of cognition.

The pursuit of real-time analysis and decision-making in sports games is destined to make a leap from perception to cognition. From the manual perspective, the referee needs to see the foul action and determine whether it is a violation and the commentator needs to pay attention to the game scene and think about the interpretation method. From the artificial intelligence perspective, the introduction of AI referee, AI commentator, and other identities is essentially an attempt to simulate the complete path from perceptual input to cognitive output. Therefore, exploring the migration path of the large language model from perception to cognition of sports scenes can not only expand its application boundary in the non-verbal intelligence field and promote the leap of the artificial intelligence from static text to dynamic scenes, but also provide intelligent technical and tactical analysis for competitive sports and bring a new watching experience for the broadcasting of sports events.

Following the main line of "from perception to cognition", this paper focuses on the application of large language models in a variety of ball games and systematically reviews the research progress at two levels. At the perception level, the ability of LLM and video coding technology to obtain sports dynamic data and action recognition is discussed. At the cognition level, the application of LLM in the tasks of tactical analysis, decision understanding, and action prediction are analyzed. Based on this, this paper summarizes the limitations of the existing technology, and outlines the future direction of the multi-modal model in the field of sports intelligence analysis.

2. Perception Level: Representation and Identification of Sports Knowledge and Scene

2.1 Sports Knowledge Coding Based on Text

Sports text is an important carrier of sports knowledge. Transforming unstructured information into structured knowledge that can be processed by a large language model is the basis of a perceptual task. How to encode discrete action types and continuous spatiotemporal information is the key link in the modeling of game event sequence. At present, the coding strategy for the event sequence of sports data mainly refers to the idea of NLP, which regards a series of competition events as sentences and each event as a word. In this framework, the core of coding is to build an event dictionary. Mendes Neves et al. Proposed an ordinal coding method for the small vocabulary of sports time series data, mapped the numerical characteristics to the front end of the encoder, coded the action type variables in order according to the frequency, and set special marks such as <PERIOD_OVER> to indicate the conversion of the competition stage. This coding method effectively improves the computational efficiency on the premise of ensuring the integrity of the sequence [3].

2.2 Video-based Understanding of Sports Events

The coding of plain text can provide the static common sense of sports for the model, while the description of the video is the dynamic entrance to obtain the information of the game. In order to systematically evaluate the model's ability to understand sports video, researchers need to model events, layer tasks, and extract features. Taking tennis as an example, the short video of less than 10 seconds may contain intensive information such as the player's choice of batting mode, the trajectory of movement, and the position of the tennis landing point. Researchers modeled the information as a time-sequential round sequence composed of continuous batting events. Through the TennisTV benchmark set, they introduced atomic events such as action recognition (AR) and hit direction (HD), sequences such as technology recognition (TI) and technology preference (TP), and built a modeling idea from atomic events to complete sequences, providing a more robust test platform for tennis videos with high information content [4]. In terms of feature extraction, the hybrid vision-language deep learning method shows good application potential. Taking ChatMatch, an intelligent analysis system for racket sports, as an example, its dedicated visual encoder uses a three-step process to extract spatial location information: semantic segmentation to identify site pixels, calculation of envelope quadrangle, and mapping 2D graphics to 3D space through perspective transformation. On this basis, it realizes multi-task recognition of sports mobilization behavior. In the decoding process, the visual information is transformed into a format that LLM can understand. This design takes into account the narrative understanding and numerical computing ability of LLM [5].

Therefore, the video based sports time understanding has gradually formed a hierarchical modeling idea, and realized feature extraction and knowledge transformation through a hybrid visual-language architecture, which laid the foundation for the LLM from perception and cognition.

2.3 Datasets for Sports Perception

The realization of the sports perception task is inseparable from high-quality data support. In terms of datasets, Xia et al sorted the sports datasets into three categories: linguistic, multi-modal, and convertible. The language type is text-based, the multi-modal type integrates video, text, audio, and other content, and the convertible type can upgrade the single-mode datasets to multi-modal resources [6].

Table 1 sorts out different types of datasets in the field of sports [6-8]. It can be seen that the current datasets construction in the field of sports is showing a multi-modal and multi-level development trend. Multi-modality is embodied in the coexistence of text, image, video and other data types. Multi-level is reflected in the fact that data cover multiple cognition levels, from basic action recognition to tactical analysis. The diversified data provides a solid data foundation and strong support for the large language model to realize the rapid development from the simple recognition of data at the perception level to the deep understanding and reasoning at the cognition level.

Table 1: Overview of sports-related datasets for intelligent analysis.

Datasets	Type	Scale & key features	Representative task/ usage
SportQA	Language-based	Three tiers of MCQs [7]	Assessing sports knowledge and reasoning ability
Sports-QA	Multi-modal (video+text)	94,000 video-Q&A pairs. Includes descriptive, temporal causal and counterfactual reasoning tasks [6]	Multi-modal event understanding and reasoning
SPORTU-video	Multi-modal (slow-motion videos)	Slow-motion videos across multiple ball sports [8]	Fine-grained action understanding and multi-modal reasoning
X-VARS	Multi-modal	Football-specific multi-modal datasets with expert annotations [6]	Tactical analysis and football event understanding
EIGD	Multi-modal	Football match videos with structured annotations [6]	Event recognition and tactical inference
FineGym	Single-modal to multi-modal	Gymnastics videos with hierarchical annotations; can be extended into multi-modal datasets via added descriptions [7]	Complex action recognition and sequence modeling
MultiSports	Single-modal to multi-modal	Multi-sport action videos; convertible to multi-modal resources through additional annotation [7]	Cross-sport action analysis and scenario understanding

3. Cognition Level: Reasoning and Prediction of the Sports Scene Based on Perception

3.1 Tactical Analysis and Decision Modeling

Transforming the sequence of events into a computable decision-making process is the basis for realizing the cognitive analysis of sports scenes. Referring to the sequence modeling idea of LLM, Mendes Neve et al Proposed the Large Events Model (LEM), which regards a series of events in a football match as sentences. Each event is like a word in a sentence. A model can predict what will happen next, and can also see the tactical intention and evaluate the performance of players [3]. They labeled each event with 11 features, such as the type of action on the court, the specific location on the court, and put this information into an event dictionary containing 140 words [3].

In multi-modal scenarios, the ChatMatch system developed by Zhang et al. Realizes tactical Q&A and situation analysis through LLM agents. The system is designed with four modules: task recognizer, coaching agent, statistical agent, and video manager, which respond to users' professional consultation through an automatic cooperation mechanism. Experiments showed that the success rate of a statistical agent on simple problems is 100%, and the success rate of moderately difficult problems is 90.9% [5].

Therefore, modeling the game as a computable sequence or through multi-agent cooperation can effectively realize tactical analysis and lay the foundation for decision-making in the game.

3.2 Decision Understanding and Player Evaluation

How to quantify every decision of players on the court is one of the core issues of sports cognitive analysis. The VAEP framework provides a solution: define the value of each action as the increased value of the team's scoring probability minus the increased value of loss probability after execution, so as to quantify the contribution of passing, shooting, dribbling, and other decisions [3]. The situational expected goal map further refined the assessment of shooting decisions. By simulating millions of shots, it calculated the scoring probability of each position in different game situations, reflecting whether the player's choice in a specific situation was reasonable [3].

Players' decisions are not only influenced by tactics, but also closely related to their psychological state. The IZOF model points out that every athlete has an optimal emotional intensity range, and exceeding this range may lead to the decline of decision quality and performance fluctuation [9]. Based on the interview text of NBA players before the game, Oved et al Found that language signals can predict the performance deviation of players in subsequent games, and the text model is better than a pure statistical model in terms of multiple tasks, indicating that the psychological state information related to subsequent decision-making is implicit in players' words [9].

The judgment of external experts can also provide a supplement for decision evaluation. Beal et al. Combined the game forward-looking text written by media experts with statistical data, and through integrated learning and fusion of text features, statistical models, and gambling company odds, achieved an accuracy of 63.18% in the prediction of Premier League game results, which was 6.9% higher than the traditional method. The model can identify 38.9% of the unexpected results, and these unconventional results are often related to the team's decision-making under pressure [10].

Therefore, from the on-site action value to the off-site language signal, and then to the fusion of expert knowledge, multidimensional information complementation breaks through the limitations of a single model and realizes a deeper understanding of decision-making.

3.3 Game Trend and Result Prediction

At the prediction level, LEM can support multi-time scale probability modeling. In the short time scale, the model can calculate the scoring probability and construct the game momentum index based on the real-time state after the event, so as to capture the dynamic evolution in the process of the game. In the long-term scale, the prediction of key results such as victory, draw, and defeat, and the number of goals can be achieved through the simulation of the whole game [3]. The benchmark datasets also provide structural support for the evaluation of prediction ability. For instance, SportQA divides the evaluation difficulty into basic facts (level-1), rules and tactics (level-2), and complex scenario reasoning (level-3). Level-3 has higher

requirements for cross event correlation, scene understanding, and game situation judgment, which can be used to test the comprehensive cognitive performance of the model in trend inference [7]. The existing results show that the model can show a certain potential of semantic understanding and trend prediction in the multi-level reasoning task, and provide an effective technical path for the event analysis task [7].

In finer-grained sports, the prediction task is often accompanied by complex trajectory inference and multi-step decision-making. Taking TennisTV as an example, the reinforcement learning reasoning model is significantly better than the non-reasoning baseline in the prediction scenario that requires multi-step aggregation and trajectory reasoning. However, in terms of forecasting tasks that emphasize time sequence positioning, such models do not show stable advantages, showing that time sequence perception is still the key bottleneck affecting the prediction effect [4]. Besides, the granularity of time series modeling needs to match the prediction task level. For example, 32 frame sampling has the best effect in batting layer prediction, while too high frame rate may weaken the model's overall grasp of turn layer trend [4].

Overall, the current model still has limitations in the key links of game trend capture and result prediction. It is important to design appropriate modeling strategies for different prediction tasks. Fine-grained time series can help capture local changes, while macro trend prediction needs to avoid information redundancy and maintain a stable understanding of the overall situation.

4. Challenges of Large Language Model in Sports Intelligence Analysis

4.1 Data Level

High quality datasets are the basis for the application of training large language model in the field of sports. However, at present, data resources are obviously insufficient. The relative scarcity of multi-modal data makes the datasets integrating video, text, and audio far less than the language datasets [6]. The inherent defects of the event data, such as the lack of key information about running without the ball in football matches (accounting for 98%), and the scale of the public datasets obtained from the Internet, which is difficult to support the training of the advanced architecture, make it difficult to grasp the error of the model itself [3].

4.2 Model Level

The current model still has great defects in the task of sports understanding, and cannot well simulate the thinking process of human brain. First, the reasoning and decision-making ability of the model in complex scenes still needs to be improved. The accuracy rate of GPT-4 in SportQA level-3 is only 47.14%, and the gap with human experts is as high as 45% [7]. Secondly, the reasoning of the model is insufficient. Even if the sub questions can be answered correctly, the error rate of the main question is still as high as 24%-44% [11]. Moreover, the temporal perception ability is weak, and the reasoning model has not been steadily improved in the task of dealing with the temporal perception ability [4].

5. Future Outlook

Looking forward to the future, the development of LLM in the field of sports intelligence analysis should be promoted along two main lines of data construction and model evolution. Since the scarcity of multi-modal resources is the main bottleneck restricting the development of the model, the upgrading of the convertible datasets can be promoted in the future, and the existing single-mode video data can be transformed into multi-modal resources through post annotation. At the same time, researchers need to consider a variety of richer information dimensions such as running without the ball and player's physiological signals, so that the model can more fully understand the dynamics of the game. Additionally, the key breakthrough is to improve the temporal awareness of the model. Future post reinforcement learning training should pay more attention to time sequence understanding and alignment mechanism. In order to improve the reasoning stability of the model, it is necessary to further explore the application of advanced architectures, such as Transformer in the sequence of sports events and introduce higher-dimensional information, such as player identity and team tactics.

6. Conclusion

Based on the theme of "from perception to cognition", this paper systematically combs and summarizes the research on the integration of LLM and sports intelligence analysis. The research shows that LLM has initially formed a solid foundation in the transformation process from the perception layer to the cognition layer.

At the perception level, relying on the processing ability of sports text, sports video, and other modes, the model can effectively obtain the basic knowledge system in the sports field and the core information in the event scene. Based on this development path, researchers have further promoted the construction and expansion of multi-type and multi-modal sports datasets, thus laying the data resources and technical framework for higher-level sports information intelligent analysis.

At the cognition level, LLM shows a trend of evolution from shallow to deep and multi-modal synergy. From data collection, modeling, and training, to model understanding and inference, and then to the prediction of key competition trends, the related technical links are gradually forming a more complete cognitive process, which continuously improves the depth and accuracy of sports intelligence analysis. However, due to the inherent complexity of sports scenes, the diversity of data structures, and the dynamic changes of application requirements, this cross field still faces systematic challenges in the standardization of datasets construction, the scalability of model application, and the stability of reasoning.

Looking ahead, research work urgently needs to promote the iterative upgrading of high-quality, convertible, and reusable datasets at the data level to meet the needs of multi-modal fusion and complex scene perception. In terms of model evolution, it is necessary to further strengthen the temporal perception ability, situational understanding ability, and cross-modal reasoning stability of LLM. Through the above efforts, it is expected to gradually realize the substantial transformation of LLM from an auxiliary tool to a "cognitive partner" in the field of sports, and provide new impetus for the theoretical expansion and practical innovation of sports intelligent analysis.

References

[1] Jiang L., Tang H., & Chen Y.(2024). A review of natural language processing based on Transformer models. Modern Computer, 30(14), 31-35.

[2] Zhao T., Xu M., & Chen A.(2025). A review of natural language processing research. Journal of Xinjiang Normal University (Philosophy and Social Sciences Edition), 46(02),89-111+2. https://doi.org/10.14100/j.cnki.65-1039/g4.20230804.001.

[3] Mendes-Neves, T., Meireles, L., & Mendes-Moreira, J. (2024). Forecasting events in soccer matches through language. arXiv preprint arXiv:2402.06820.

[4] Bao, Z., & Zhang, L. (2025). TennisTV: Do Multimodal Large Language Models Understand Tennis Rallies?. arXiv preprint arXiv:2509.15602.

[5] Zhang, J., Han, D., Han, S., Li, H., Lam, W. K., & Zhang, M. (2025). ChatMatch: Exploring the potential of hybrid vision–language deep learning approach for the intelligent analysis and inference of racket sports. Computer Speech & Language, 89, 101694.

[6] Xia, H., Yang, Z., Zhao, Y., Wang, Y., Li, J., Tracy, R., ... & Shen, W. (2024). Language and multimodal models in sports: A survey of datasets and applications. arXiv preprint arXiv:2406.12252.

[7] Xia, H., Yang, Z., Wang, Y., Tracy, R., Zhao, Y., Huang, D., ... & Shen, W. (2024, June). Sportqa: A benchmark for sports understanding in large language models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 5061-5081).

[8] Xia, H., Yang, Z., Zou, J., Tracy, R., Wang, Y., Lu, C., ... & Chen, H. (2024). Sportu: A comprehensive sports understanding benchmark for multimodal large language models. arXiv preprint arXiv:2410.08474.

[9] Oved, N., Feder, A., & Reichart, R. (2020). Predicting in-game actions from interviews of NBA players. Computational Linguistics, 46(3), 667-712.

[10] Beal, R., Middleton, S. E., Norman, T. J., & Ramchurn, S. D. (2021, May). Combining machine learning and human experts to predict match outcomes in football: A baseline model. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 17, pp. 15447-15451).

[11] Yang, Z., Xia, H., Li, J., Chen, Z., Zhu, Z., & Shen, W. (2025). Sports intelligence: Assessing the sports understanding capabilities of language models through question answering from text to video. Electronics, 14(3), 461.

Funding

This research received no external funding.

Conflicts of Interest

The authors declare no conflict of interest.

Acknowledgment

This paper is an output of the science project.

Copyrights