Generating Tennis Action Instruction Based on a Large Language Model

Zhiyuan Wang[1], Zhiqian Xia[1], Guangyao Zhao[1], Kaiyue Lu[2], Haifeng Xia[1], Siyu Xia[1]
[1]Advanced Ocean Institute, Southeast University, China
[2]Nantong Productivity Promotion Center, China
{zhiy.wang, zhi.xia, gyzhao, hfxia, xsy}@seu.edu.cn, kaiyue13348096653@163.com

Abstract

*As tennis continues to gain global popularity, there is a growing need for precise, scalable, and standardized training solutions. However, the evaluation and instruction of tennis techniques remain largely dependent on subjective assessments by experienced coaches, limiting consistency and automation. In this paper, we present **TAGS** (Tennis Action Guidance System), a unified framework that leverages large language models to generate expert-level evaluations and actionable feedback for tennis performance. Given a tennis video and a user-defined query, TAGS first extracts 3D skeletal keypoints, encodes the motion into a structured representation, and employs prompt-based LLMs to assess performance across five critical dimensions: stability, coordination, power generation, technical execution, and rhythm. To support this process, we introduce (1) a pose–language unified representation that aligns 3D motion data with natural language understanding, (2) structured prompt templates grounded in professional coaching expertise, and (3) a multi-dimensional evaluation protocol incorporating both human and automated assessments. In addition, we contribute the first large-scale dataset for tennis action evaluation, comprising 3,000 video–query–feedback triplets authored by certified coaches. Extensive experiments demonstrate that TAGS consistently outperforms existing baselines in delivering accurate, domain-specific, and actionable feedback. This work marks a step forward in developing interpretable, human-aligned systems for real-world sports action assessment and intelligent coaching.*

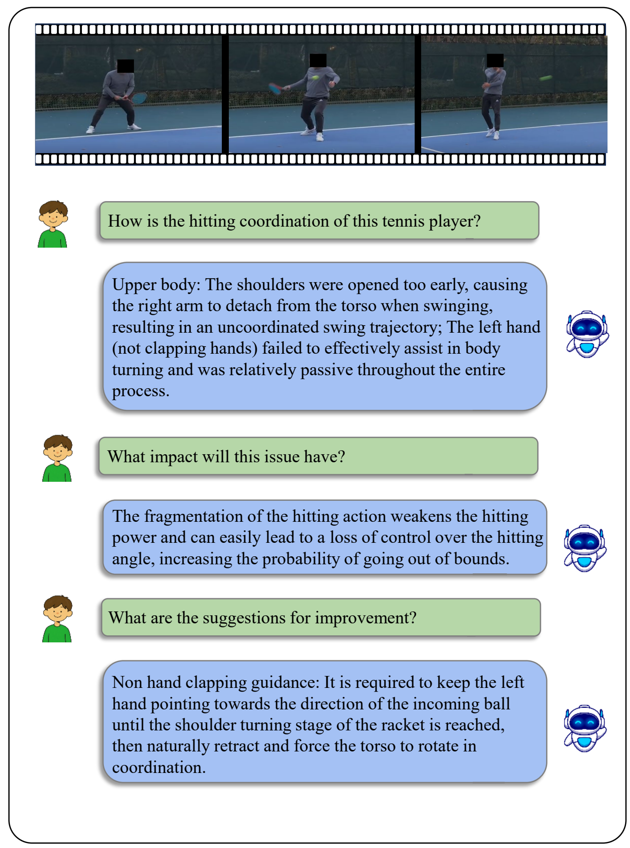

Figure 1. An example TAGS system response for tennis action evaluation, given a video and user query. This sample highlights hitting coordination analysis—issues, impacts, and improvement suggestions—while the full output also addresses stability, power, technique, and rhythm.

1. Introduction

In the domain of intelligent sports, tennis serves as a representative task due to its strong dependence on precise body coordination and technical execution. Traditional methods for evaluating tennis actions predominantly rely on the subjective observation and verbal feedback of experienced coaches. However, such approaches suffer from inconsistent evaluation criteria, high inter-observer variability, and limited scalability, thereby making it particularly difficult to deliver high-quality, personalized guidance to amateur players.

While some prior studies have attempted to apply deep learning techniques to the recognition and classification of tennis actions [2, 21, 28, 44], most of these efforts are constrained to identifying what the action is, rather than evaluating how well it is performed or how it could be improved. More recently, certain approaches [11, 13] have explored converting single-frame 3D poses or short motion sequences into textual descriptions for the purpose of coarse-grained behavioral understanding or classification. Nevertheless, these methods typically demonstrate limited multimodal modeling capacity, lack fine-grained analysis, and fail to provide actionable, instructional feedback. Crucially, they do not establish a closed-loop system that connects perception with expert-level guidance.

With the rapid advancement of Artificial Intelligence, particularly Large Language Models (LLMs) [6, 9, 12, 17, 19, 37, 41], significant progress has been made in the understanding and generation of multimodal information. These models have shown great promise in handling visual, temporal, and structural data such as images [3, 26], videos [4, 25, 45], and motion sequences [8, 20, 46]. In complex and highly dynamic sports like tennis, this technological trend offers new opportunities for automated evaluation and personalized instruction.

However, technical movements in tennis—such as swinging, jumping, and torso rotation—often occur within very short time windows and involve subtle temporal dynamics. If a model fails to accurately capture these fine-grained, time-sensitive features, the resulting interpretation is likely to be incomplete or misleading, leading to unreliable and impractical feedback. This challenge highlights the inadequacy of current LLMs in capturing complex temporal dynamics inherent in human motion, and exposes their limited temporal sensitivity and reasoning capability in high-speed action modeling. Therefore, enhancing the temporal perception of LLMs for fast-paced motion sequences and effectively translating this understanding into structured, professional, and instructive feedback is a critical challenge in building real-world action assessment and coaching systems.

To address these issues, we propose a unified LLM-based framework for tennis action assessment and guidance. Given a tennis video and a user-defined query, our system first extracts 3D skeletal keypoints and encodes the motion into a structured representation. The encoded data is then fed into a large language model, which generates comprehensive feedback covering five core dimensions: stability, coordination, power generation, technical execution, and rhythm/continuity. The output highlights key weaknesses, their potential impacts, and corresponding improvement suggestions (see Figure 1).

To support this framework, we introduce three key mechanisms. First, we propose a pose–language unified representation, in which multi-frame 3D skeletal sequences are encoded into a structured format that aligns with natural language input, allowing LLMs to interpret tennis actions within a unified semantic space. Second, we employ structured prompt engineering using expert-informed templates to guide the generation process, ensuring that the model outputs are coherent, professional, and actionable. Third, we establish a multi-dimensional evaluation protocol that incorporates both expert human ratings and automatic metrics to assess the generated feedback across four critical dimensions: accuracy, expertise, readability, and usability.

Additionally, we construct the first expert-level tennis action assessment dataset, containing over 3,000 video–question–feedback triplets written by professional coaches. This fills a critical gap in multimodal datasets that combine fine-grained actions with corresponding expert commentary.

Our main contributions are summarized as follows: (1) We propose **TAGS**, a unified framework for tennis action understanding and instructional feedback, which establishes a closed-loop system that maps multimodal input to expert-level guidance. (2) We construct the first large-scale and structured dataset for tennis action assessment, featuring high-quality textual annotations and well-aligned question-answer pairs. (3) We develop a comprehensive evaluation methodology, including prompt design analysis, benchmarking against existing multimodal LLMs, and a human study comparing the performance of the TAGS system with that of professional coaches.

2. Related work

Multimodal Large Language Models. The rapid advancement of large language models (LLMs), such as GPT, Qwen [43], and LLaMA, has catalyzed progress in unified modeling across modalities including vision, text, audio, and video. In the domain of motion understanding, researchers have explored the use of LLMs to process 3D poses, image frames, and video sequences by encoding them into representations compatible with language inputs. Representative works such as UniPose [1, 24], MotionGPT [20], and ChatPose [13] adopt token-level discretization of skeletal keypoints, enabling LLMs to be applied to pose generation, understanding, and editing tasks. While these multimodal LLMs demonstrate strong cross-modal reasoning abilities in general contexts, two limitations persist in the context of sports: (1) a lack of modeling capacity for fast, fine-grained motion sequences, and (2) an inability to produce domain-specific and actionable guidance, often resulting in abstract or generic descriptions.

Video-Based Action Understanding. Video-based action understanding remains a central topic in computer vision, encompassing tasks such as action recognition, temporal localization, and behavior segmentation. Earlier approaches relied on spatiotemporal convolutional networks (e.g., C3D) or optical flow-based methods [38, 40]. More recently, transformer-based architectures such as TimeSformer [5], VideoMAE [36], and STCFormer [34] have emerged, offering improved temporal dependency modeling for long-range sequences. In sports contexts, several studies have applied 3D convolution or skeleton-based networks to recognize and score specific movements [2, 21]. However, these works primarily address the question of "what the action is," rather than "whether the action is correct" or "how it can be improved." Some efforts have explored action quality assessment via regression-based scoring [15, 31, 42], but often lack interpretable, language-based feedback, limiting their applicability to real coaching or instructional scenarios.

Action Assessment and Instructional Feedback. Action assessment tasks go beyond recognition, requiring models to evaluate the quality of a given movement across multiple dimensions and provide constructive feedback. This problem combines knowledge of biomechanics, motion semantics, and natural language generation, making it significantly more complex than classification or generation alone. Some recent studies [10, 11, 14, 22] attempt to construct pose-pair-to-text systems that describe differences between two actions, but these typically focus on generic or daily behaviors rather than domain-specific athletic performance. In practice, motion evaluation in sports still relies heavily on expert coaches, and automated systems remain underdeveloped. Existing models often fail to capture important aspects such as temporal continuity, rhythm, and inter-limb coordination. Moreover, generated text tends to lack the precision, professionalism, and practicality required in real-world settings. Therefore, building systems that integrate high-quality motion representation with structured, context-aware language feedback, especially in domains such as power mechanics, timing, and stability, remains a critical and underexplored challenge in intelligent sports training.

3. Dataset

The performance of supervised models is fundamentally dependent on the quality and quantity of the available annotated datasets. Multimodal datasets, in particular, present greater challenges in terms of management and annotation due to the complexity of aligning diverse modalities. As a result, they are often smaller in scale compared to unimodal datasets. Table 1 presents a comparative summary of existing multimodal datasets in the literature, with a focus on those comprising text-image and text-video pairs.

To facilitate fine-grained, interpretable, and structured evaluation of tennis techniques, we introduce a novel expert-annotated dataset for tennis action assessment. Each sample in the dataset consists of a short video clip, a corresponding question about a specific tennis movement, and a detailed answer that provides both assessment and guidance. The answers are systematically organized into five evaluation dimensions—**stability**, **coordination**, **kinetic chain**, **technical compliance**, and **rhythm**—and are further decomposed into three analytical components: **weaknesses**, **impacts**, and **recommendations**. This hierarchical annotation structure enables the dataset to support multimodal models that integrate both language and action understanding tasks.

3.1. Dataset Composition

To support fine-grained analysis and generation of tennis action assessments, we construct a novel dataset consisting of 3,000 tennis action videos, each paired with a detailed natural language evaluation. To the best of our knowledge, this is the first large-scale dataset specifically designed for tennis action assessment, bridging the gap between raw motion data and professional coaching-level feedback. Notably, each evaluation exceeds 800 words, providing comprehensive and in-depth feedback that enables athletes to better understand their technical shortcomings and clear directions for improvement.

Each data sample includes:

- **Tennis action video:** A single clip showcasing a specific type of tennis action (e.g., backhand, volley, serve).
- **3D skeleton sequence:** A frame-wise sequence of 17 keypoints extracted from the video, representing essential human joints (e.g., hips, spine, shoulders, knees).
- **Natural language evaluation:** A paragraph-level structured assessment written in fluent, instructional language, simulating how a human coach would comment on the motion.

The action evaluation is organized around five core performance dimensions that align with coaching practice:

- **Stability:** e.g., body balance, center of mass control.
- **Coordination:** e.g., interaction between limbs and torso.
- **Power Generation Mechanism:** e.g., kinetic chain, timing.
- **Technical Specification:** e.g., preparation, swing path, follow-through.
- **Rhythm and Continuity:** e.g., flow, timing disruptions.

For each dimension, the evaluation typically includes:

- **Observed Issues:** A clear description of the technical or biomechanical problem.
- **Potential Impact:** The effect of the issue on performance consistency, power, or accuracy.
- **Targeted Suggestions:** Actionable improvement strategies rooted in professional training techniques.

Additionally, each evaluation concludes with a core improvement strategy section, summarizing key training recommendations that span multiple performance aspects. These strategies are designed to be practical and implementable, such as "balance-board shadow swings" or "multi-ball interception drills," making the dataset highly valuable for real-world coaching scenarios and AI learning alike.

Dataset	**Type**	**Instances**	**SK**	**Context**	**VQA**	**Captioning**	**AE**
Conceptual 12M [7]	Images	12M	–	Various	–	✓	–
LAION-5B [33]	Images	3B	–	Various	–	✓	–
LLaVA dataset [27]	Images	158k	–	Various	–	✓	–
MovieQA [35]	Videos	408	–	Movies	✓	✓	–
TVQA [23]	Videos	21k	–	TV shows	✓	✓	–
Video Instruction Dataset [29]	Videos	100k	–	Various	✓	✓	–
HowTo100M [30]	Videos	136M	–	Youtube	–	✓	–
TennisPose [39]	Videos	2k	2D	Tennis	–	✓	–
THETIS [16]	Videos	8374	3D	Tennis	–	✓	–
Tennis Action-GE (Ours)	Videos	3k	3D	Tennis Action Evaluation	✓	✓	✓

Table 1. Comparative overview of relevant datasets. Tennis Action-GE contains high-quality tennis action evaluations annotated by 20 experienced coaches. It is the only dataset focused on expert-level evaluation of tennis technique. SK denotes types of skeletal keypoints, AE refers to action evaluation, and VQA stands for visual question answering.

3.2. Dataset Novelty and Strengths

This dataset has several distinctive features:

- **First-of-its-kind:** No existing public datasets provide structured, multi-aspect assessments of tennis actions in natural language form.
- **Expert-aligned instruction:** Each evaluation mirrors the analysis logic of real tennis coaches, making it highly interpretable and pedagogically relevant.
- **Action-level granularity:** Assessments are not generic; they are tailored to specific stroke types (e.g., volley, forehand, slice), with contextual feedback.
- **Rich semantic structure:** Evaluations include multi-paragraph analysis with subheadings and logical progression (*problem* $\rightarrow$ *impact* $\rightarrow$ *advice*), facilitating both supervised learning and NLG tasks.
- **Application-ready:** The dataset is well-suited for training and evaluating LLMs on tasks such as motion captioning, action diagnosis, performance coaching, and instructional feedback generation.

3.3. Dataset Construction

To construct a high-quality dataset for tennis action assessment, we propose a hybrid pipeline that combines LLM-based analysis with expert refinement. The entire process includes three main stages: automatic generation, expert revision, and final validation.

Automatic Evaluation Generation. In the first stage, we utilize a large language model (DeepSeek [18]) to generate preliminary action assessments. For each sample, we provide two sequences of 3D human keypoints extracted from tennis action videos: one representing a standard demonstration (labeled as coach), and the other a non-standard performance (labeled as student). Each sequence consists of a series of frames, and each frame contains 17 keypoints representing key body parts, including the hips, spine, shoulders, elbows, knees, wrists, and ankles.

To elicit meaningful analysis from the language model, we design a domain-specific prompt that asks the model to evaluate the student's tennis movement—such as a backhand stroke—from five core perspectives: stability, coordination, power generation mechanism, technical specification, and rhythm & continuity. The prompt instructs the model to avoid numerical or coordinate-based descriptions and to provide natural language explanations, as if observing a real tennis action video.

The resulting output from the model is a structured, professional-level assessment written in natural language. Each output identifies key weaknesses, potential impacts, and specific suggestions for improvement, thereby forming the initial version of our action evaluation data.

Expert Revision and Enrichment. To ensure the quality and authenticity of the dataset, we invite 20 experienced tennis coaches to manually revise the automatically generated assessments. Each coach is asked to watch the actual tennis action videos associated with each sample and refine the evaluation text accordingly. Coaches are encouraged to freely pause, replay, or skip frames to avoid fatigue and ensure accuracy. They are given full flexibility to rewrite or restructure any part of the assessment to better match real-world coaching logic and expression.

This expert-in-the-loop revision process guarantees that the evaluations not only reflect professional judgment, but also retain high readability and instructional value for both humans and models.

Final Quality Check. After expert revision, a final validation is conducted by an additional group of 10 senior

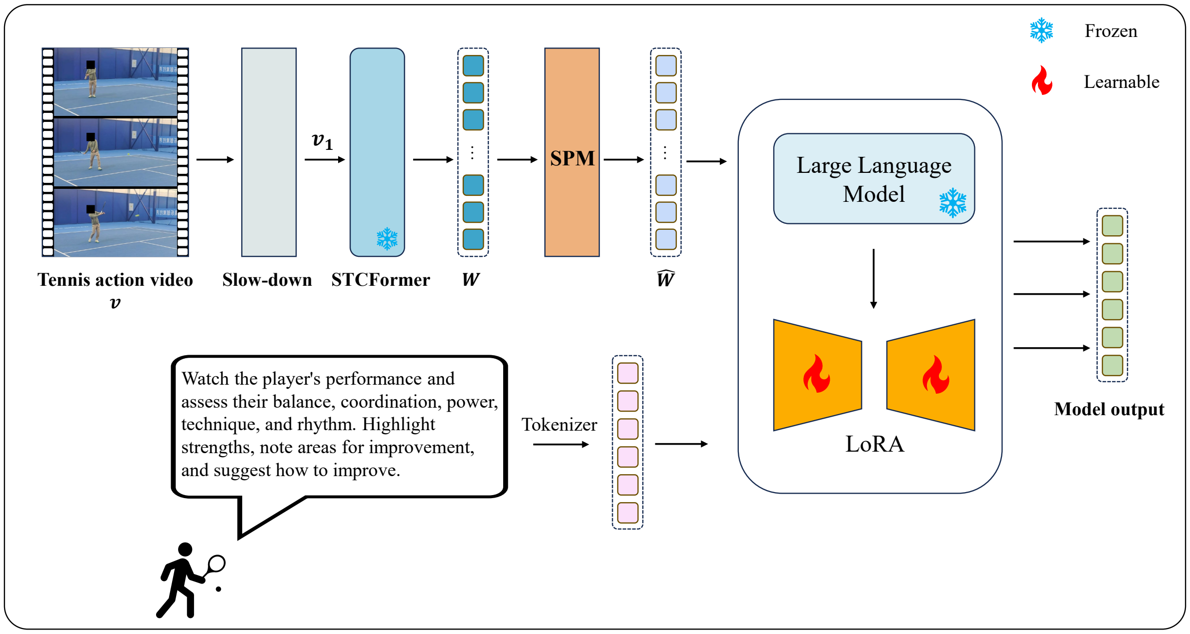

Figure 2. Architecture of TAGS. TAGS is a tennis action guidance system designed to evaluate athletic performance based on skeleton-based motion features. The system first extracts spatio-temporal skeleton sequences from input tennis videos using a pretrained STCFormer model. These features are first input to a large language model (Qwen2.5-7b), after which a linear projection aligns them to the text embedding space required by the model. The language model has been fine-tuned using LoRA to generate structured motion assessments. The prompt includes five evaluation dimensions—balance, coordination, power, technique, and rhythm—and the model outputs natural language feedback detailing weaknesses, impacts and suggestions for improvement. Frozen parameters in both the vision encoder and language model ensure stable representation learning, while LoRA updates only a small set of learnable parameters.

coaches, who cross-check all samples to eliminate factual errors, misjudged assessments, or redundant phrasing. This step ensures a consistent, reliable, and error-free dataset suitable for downstream applications in tennis action understanding, generation, and coaching feedback systems.

4. Methodology

Figure 2 illustrates the overall architecture of our proposed **Tennis Action Guidance System (TAGS)**. The system primarily consists of two key components: STC-Former [34], a model specialized in extracting 3D human skeletal keypoints from video sequences, and Qwen2.5-7B [32], a large-scale language model with 7 billion parameters designed for reasoning and evaluation.

4.1. Architecture

Given an input tennis action video $\mathbf{v} \in \mathbb{R}^{T\times H\times W\times C}$, where T, H, W, and C represent the number of frames, height, width, and channel size respectively, the video is first passed through a temporal slow-down module. This module addresses the challenge of high-speed motion typically present in athletic movements, producing a temporally expanded video $\mathbf{v}_1 \in \mathbb{R}^{S\times T\times H\times W\times C}$, where S is the slow-motion scaling factor.

The slowed-down video $\mathbf{v}_1$ is then processed by the STC-Former to extract 3D skeletal joint coordinates. For each frame, we obtain 17 anatomical keypoints represented by:

$$\mathbf{w}_t = \{(x_i, y_i, z_i)\}_{i\in\mathcal{I}} \tag{1}$$

where i indexes the anatomical keypoints including pelvis center, spine, chest, neck, head, right hip, right knee, right foot, left hip, left knee, left foot, right shoulder, right elbow, right wrist, left shoulder, left elbow, and left wrist.

Over a sequence of T frames, we denote the full skeletal sequence as:

$$\mathbf{W} = \{\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_T\}, \quad \mathbf{w}_t \in \mathbb{R}^{\mathcal{I}\times D} \tag{2}$$

where $\mathbf{W}$ denotes the sequence of skeletal poses over T frames, $\mathbf{w}_t$ is the pose at time t, $\mathcal{I}$ is the number of anatomical keypoints, and D is the spatial dimension (e.g., $D = 3$ for 3D coordinates).

Before feeding the skeletal sequence into the language model, a skeleton preprocessing module is applied to compress the data and reduce token length, while preserving essential motion characteristics. Since transformer-based models like Qwen2.5-7B have a limited context window, even a short (e.g., 3-second) video may generate a sequence exceeding the model's input capacity.

To address this, we propose a two-step preprocessing pipeline:

1. **Temporal subsampling**: Retain only one out of every five frames.
2. **Floating-point truncation**: Round each coordinate (x, y, z) to 4 decimal places.

The resulting compact representation is denoted as:

$$\hat{\mathbf{W}} = \mathrm{SPM}(\mathbf{W}, K, L) \tag{3}$$

where $\mathbf{W}$ denotes the input initial skeleton sequence, and K, L are hyperparameters controlling the SPM process, set respectively to 4 and 5.

The preprocessed skeleton sequence $\hat{\mathbf{W}}$ is embedded into a natural language prompt tailored to the tennis action evaluation task and fed into the Qwen2.5-7B model. This architecture allows TAGS to bridge the gap between low-level motion features and high-level semantic evaluation and guidance, delivering expert-level diagnostic feedback in natural language.

4.2. Fine-tuning Strategy and Prompt Design

During the fine-tuning phase of our large language model (Qwen2.5-7B), we do not directly use raw video frames paired with their corresponding action assessments as training data. Instead, we utilize the extracted 3D skeletal keypoint sequences as input, paired with expert-written tennis action evaluations as supervision. Importantly, the weights of STC-Former remain frozen throughout this process, ensuring that the model continues to produce consistent and domain-invariant representations of human motion. Only the Qwen2.5-7B model is fine-tuned, enabling it to acquire domain-specific knowledge related to tennis action understanding and evaluation. To perform parameter-efficient adaptation, we employ Low-Rank Adaptation (LoRA) during fine-tuning. This approach significantly reduces GPU memory consumption by injecting trainable rank-decomposed matrices into each layer of the language model, while keeping the original weights frozen. As a result, the model can be effectively adapted to the tennis domain with minimal computational overhead.

To enhance both the response relevance and generation efficiency of the model, we designed a task-specific prompt template tailored to the tennis action evaluation task. The prompt is constructed to guide the LLM to emulate the perspective of a professional tennis coach, and is structured as follows:

USER : <Question> <EvaluationFocus> $<\hat{\mathbf{W}}>$
Assistant :

where $<$ *Question* $>$ represents a user-posed query regarding tennis action assessment, $<$ *EvaluationFocus* $>$ outlines five key evaluation dimensions: stability, coordination, power generation mechanics, technical correctness, and rhythm and fluency. The variable $< \hat{\mathbf{W}} >$ denotes the 3D coordinates of human skeletal keypoints extracted from tennis action videos.Notably, the prompt design was developed in close collaboration with experienced tennis coaches, whose domain expertise ensured that the instructions accurately reflect real-world coaching language and focus areas. This collaboration contributed to the clarity, specificity, and practical utility of the generated feedback. TAGS is trained to generate responses in an autoregressive manner, conditioned on these prompts.

This carefully crafted prompt enhances the model's ability to produce accurate, instructive, and human-readable evaluations without relying on low-level numerical artifacts. By bridging structured motion representation and high-level coaching insight, our fine-tuning methodology effectively aligns LLM capabilities with the needs of automated sports performance analysis. In particular, our fine-tuning is complemented by carefully crafted prompt designs that explicitly guide the model to focus on key evaluation dimensions such as stability, coordination, power generation, technical execution, and rhythm. These structured prompts help constrain the model's output space and enhance the relevance, clarity, and actionability of the generated feedback.

5. Experiments

In this section, we design three experiments to comprehensively and rigorously evaluate the performance of **TAGS** on the task of tennis action assessment and guidance.

5.1. Evaluation Study

Evaluating generative tasks such as text, image, or video generation remains a substantial challenge due to their inherent subjectivity and the lack of universally accepted evaluation metrics. Conventional linguistic metrics are not well-suited for our purpose, as two textual responses can differ significantly in language yet convey essentially the same meaning.

To obtain quantitative results, we conducted a human evaluation study involving 10 professional tennis coaches, each with over 10 years of coaching experience and more than 1,000 students coached. Each coach was asked to assess 20 randomly selected short tennis action clips (3–5 seconds in length). The coaches were blind to the source of each response (i.e., whether it was generated by a human expert or by TAGS). To mitigate potential biases due to emotional or environmental factors, the coaches were allowed to complete the evaluations at their convenience.

The evaluation was conducted across four principal dimensions: Accuracy, Professionalism, Readability, and Practicality, each reflecting a critical aspect of assessment quality. Each dimension was rated on a 5-point Likert scale (5 = strongly agree, 1 = strongly disagree). Table 2 summarizes the evaluation results. Overall, the per-

Eval. Dim.	Acc.↑	Prof.↑	Read.↑	Pract.↑	Avg. Score
Referees	**4.70**	**4.85**	4.65	**4.60**	**4.70**
TAGS	4.62	4.75	**4.80**	4.58	4.69
Diff.	-0.08	-0.10	0.15	-0.02	-0.01

Table 2. **Comparison of Scores between Coaches and TAGS.**The performance of TAGS was comparable to that of human experts, with only minor differences in scores. Acc. = Accuracy, Prof. = Professionalism, Read. = Readability, Pract. = Practicality, Diff. = Score Difference, Eval. Dim. = Evaluation Dimension.

formance of TAGS was comparable to that of human experts, with only minor differences in scores. In particular, TAGS achieved higher scores in readability, likely benefiting from the strong language generation capabilities of Qwen2.5. However, slight deficiencies were observed in accuracy and professionalism, attributed to occasional hallucinations where TAGS introduced errors or referenced actions not actually present in the video. Regarding practicality, the lower scores were mainly due to TAGS providing less specific and actionable guidance compared to human coaches, who tended to provide more personalized training recommendations and clearly articulated drills.

Evaluator	Similarity	Clarity	Practicality
ChatGPT-4o	4.80	4.75	4.70
DeepSeek	4.75	4.78	4.68
Average	4.78	4.77	4.69

Table 3. **Automatic Evaluation of TAGS by LLMs.** ChatGPT-4o and DeepSeek were provided with two evaluation texts—one written by a professional coach and the other generated by TAGS. They were prompted to assess the similarity, clarity, and practicality of the TAGS-generated text relative to the coach's version. Scores range from 1 to 5.

We further conducted automatic evaluation using two advanced LLMs, ChatGPT-4o and DeepSeek, which compared human-written and TAGS-generated tennis action assessments across content relevance, clarity, and practical usefulness. As shown in Table 3, TAGS responses closely matched expert-level assessments, with scores nearing the maximum of 5.

In conclusion, TAGS demonstrates a strong capacity to understand tennis action videos and produce expert-aligned evaluations and suggestions, confirming its value for action analysis and personalized feedback.

5.2. Comparative Study

To validate TAGS's effectiveness in generating tennis action assessments and guidance, we conducted a comparative study with two leading multimodal LLMs, ChatGPT-4o and Qwen2.5-VL, recognized for strong visual understanding and language generation.

Experimental Setup: We selected 20 tennis video clips, each ranging from 3 to 5 seconds in length, covering a diverse set of tennis techniques, including forehands, backhands, volleys, and serves. Each video clip was associated with a predefined question, formatted similarly to those employed during the training phase. These input pairs were fed into the TAGS system, ChatGPT-4o, and Qwen2.5-VL, respectively, and their generated responses—comprising both action assessment and instructional suggestions—were collected for evaluation.To assess the perceived quality of these responses, we invited ten experienced tennis coaches and the corresponding students featured in the videos to jointly evaluate the outputs. For each video, the coaches independently selected the response they deemed most appropriate. Meanwhile, each student chose the response they considered best among the three evaluations generated specifically for their own tennis action video.

System	ChatGPT-4o	Qwen2.5-VL	TAGS
Coaches (%)	26.5%	17.5%	56.0%
Students (%)	25.0%	15.0%	60.0%

Table 4. **Human Preference for System Outputs.** Ten experienced coaches and the corresponding students were asked to select the most helpful response from three systems (TAGS, ChatGPT-4o, and Qwen2.5-VL) for each video. The table reports the percentage of times each system was preferred by coaches and students, respectively.

Results: Table 4 shows the results. The results demonstrate a clear preference for the TAGS system. These findings suggest that both expert evaluators and end users perceived the outputs from TAGS to be more informative, relevant, and actionable than those produced by the baseline models.This consistent preference across both groups highlights the effectiveness of TAGS in delivering domain-specific, user-aligned feedback that better supports tennis skill development.These findings suggest that although general-purpose multimodal language models possess strong generative capabilities, they often fall short in professional application scenarios such as sports coaching, where precise motion interpretation and actionable guidance are essential. In contrast, TAGS demonstrates clear advantages in tennis-specific analysis and instruction, producing feedback that is both contextually appropriate and practically implementable.

model version	prompt type	fine-tuning status	Accuracy↑	Professionalism↑	Readability↑	Practicality↑
Qwen2.5-7b	simple prompt	non-fine-tuned	4.65	4.10	4.30	4.05
Qwen2.5-7b	templated prompt	non-fine-tuned	4.66	4.22	4.60	4.25
TAGS	simple prompt	fine-tuned	4.68	4.80	4.69	4.46
TAGS	templated prompt	fine-tuned	**4.70**	**4.85**	**4.76**	**4.62**

Table 5. **Ablation results across model versions, prompt types, and fine-tuning status.** This table shows average scores on accuracy, professionalism, readability, and practicality for various setups. We compare two models (Qwen2.5-7B and our fine-tuned TAGS) with simple and templated prompts, both fine-tuned and non-fine-tuned. Results indicate that prompt design and fine-tuning improve response quality, with the TAGS model using templated prompts achieving the highest scores across all metrics.

5.3. Ablation Study

Effect of Prompt Design. To investigate the influence of prompt engineering on the quality of action assessment generation, we designed an ablation experiment in which the model architecture and input 3D skeletal keypoint sequences were held constant, and only the prompt formulation was varied. Specifically, we compared the performance of our proposed structured prompt template with that of a generic natural language prompt.

The structured prompt explicitly instructed the model to act from the perspective of a professional tennis coach and provided detailed evaluation dimensions, including: **stability** (e.g., balance and control of the center of mass), **coordination** (e.g., interaction between upper and lower body), **power generation** (e.g., continuity of kinetic chain and timing of force application), **technical correctness** (e.g., preparation posture, swing path, and follow-through), and **rhythm and fluency** (e.g., smoothness and avoidance of redundant motions). It also required the model to describe the motion as if observing a live-action video rather than analyzing numerical data, discouraging the use of coordinates, angles, or frame counts in the output.

In contrast, the generic prompt simply instructed the model to evaluate a tennis action based on a list of 3D keypoints, without further elaboration or guidance. To evaluate the responses generated under each prompt condition, we invited experienced human coaches to serve as evaluators. The generated outputs were scored across four criteria: accuracy, professionalism, readability, and practicality, using a 5-point Likert scale (1 = poor, 5 = excellent).

As shown in Table 5, the structured prompt substantially outperformed the generic prompt in all dimensions, with particularly notable improvements in readability and practicality. These results highlight the importance of clear, task-specific prompt design in guiding large language models to produce more coherent, relevant, and domain-appropriate outputs in specialized applications.

Fine-tuned vs. Pre-trained Models. To further assess whether fine-tuning on domain-specific data improves model performance in the tennis action assessment task, we compared two model configurations: (1) the base Qwen2.5 model without fine-tuning, and (2) the fine-tuned version used in our TAGS system. Both models were provided with the same 3D skeletal sequences and the same structured prompt template.

Table 5 shows that the fine-tuned model achieved higher scores across all dimensions, with the most substantial gains observed in professionalism and practicality. These results demonstrate that fine-tuning on task-relevant motion data significantly enhances the model's ability to understand tennis-specific biomechanics and generate more expert-level, actionable instructional feedback.

6. Conclusion

This paper proposes a unified LLM-based framework for the assessment and guidance of tennis action. Given a video and a user query, the proposed method generates structured, multidimensional feedback covering both key weaknesses and improvement suggestions. Specifically, this paper designs a pose language representation, expert-guided prompt templates, and a systematic evaluation protocol. In addition, this paper builds the first expert-annotated dataset for tennis action feedback. The experimental results show that the method provides accurate, readable, and actionable guidance. This work offers a step toward real-world intelligent sports instruction.

7. Acknowledgements

This work was supported by Key Laboratory of Intelligent Processing Technology for Digital Music (Zhejiang Conservatory of Music), Ministry of Culture and Tourism under Grant 2023DMKLB003, the Big Data Computing Center of Southeast University, and Research Fund for Advanced Ocean Institute of Southeast University, Nantong (No. GP20010002).

References

[1] Bruno Artacho and Andreas Savakis. Unipose: Unified human pose estimation in single images and videos. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 7035–7044, 2020. 2

[2] Hongzhou Bai. Tennis teaching assistance model based on double chain shared unsupervised action recognition algorithm. *International Journal of Cognitive Computing in Engineering*, 6:21–31, 2025. 1, 3

[3] Jinze Bai, Shuai Bai, Shusheng Yang, Shijie Wang, Sinan Tan, Peng Wang, Junyang Lin, Chang Zhou, and Jingren Zhou. Qwen-vl: A frontier large vision-language model with versatile abilities. *arXiv preprint arXiv:2308.12966*, 1(2):3, 2023. 2

[4] Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibo Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, et al. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*, 2025. 2

[5] Gedas Bertasius, Heng Wang, and Lorenzo Torresani. Is space-time attention all you need for video understanding? In *ICML*, page 4, 2021. 2

[6] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020. 2

[7] Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. Conceptual 12m: Pushing web-scale image-text pretraining to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 3558–3568, 2021. 4

[8] Ling-Hao Chen, Shunlin Lu, Ailing Zeng, Hao Zhang, Benyou Wang, Ruimao Zhang, and Lei Zhang. Motionllm: Understanding human behaviors from human motions and videos. *arXiv preprint arXiv:2405.20340*, 2024. 2

[9] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *Journal of Machine Learning Research*, 24(240):1–113, 2023. 2

[10] Ginger Delmas, Philippe Weinzaepfel, Francesc Moreno-Noguer, and Grégory Rogez. Posefix: correcting 3d human poses with natural language. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 15018–15028, 2023. 3

[11] Ginger Delmas, Philippe Weinzaepfel, Thomas Lucas, Francesc Moreno-Noguer, and Grégory Rogez. Posescript: Linking 3d human poses and natural language. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 47(7):5146–5159, 2025. 2, 3

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019. 2

[13] Yao Feng, Jing Lin, Sai Kumar Dwivedi, Yu Sun, Priyanka Patel, and Michael J. Black. Chatpose: Chatting about 3d human pose. In *2024 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2093–2103, 2024. 2

[14] Mihai Fieraru, Mihai Zanfir, Silviu Cristian Pirlea, Vlad Olaru, and Cristian Sminchisescu. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9919–9928, 2021. 3

[15] Kumie Gedamu, Yanli Ji, Yang Yang, Jie Shao, and Heng Tao Shen. Fine-grained spatio-temporal parsing network for action quality assessment. *IEEE Transactions on Image Processing*, 32:6386–6400, 2023. 3

[16] Sofia Gourgari, Georgios Goudelis, Konstantinos Karpouzis, and Stefanos Kollias. Thetis: Three dimensional tennis shots a human action dataset. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 676–681, 2013. 4

[17] Suriya Gunasekar, Yi Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Mojan Javaheripi, Piero Kauffmann, Gustavo de Rosa, Olli Saarikivi, Adil Salim, Shital Shah, Harkirat Singh Behl, Xin Wang, Sébastien Bubeck, Ronen Eldan, Adam Tauman Kalai, Yin Tat Lee, and Yuanzhi Li. Textbooks are all you need, 2023. 2

[18] Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, et al. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *arXiv preprint arXiv:2501.12948*, 2025. 4

[19] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. 2

[20] Biao Jiang, Xin Chen, Wen Liu, Jingyi Yu, Gang Yu, and Tao Chen. Motiongpt: Human motion as a foreign language. *Advances in Neural Information Processing Systems*, 36:20067–20079, 2023. 2

[21] Li Ke. Automated detection of tennis training movements: Leveraging deep learning and multiscale image decomposition. *International Journal of High Speed Electronics and Systems*, page 2540594, 2025. 1, 3

[22] Hyounghun Kim, Abhay Zala, Graham Burri, and Mohit Bansal. Fixmypose: Pose correctional captioning and retrieval. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 13161–13170, 2021. 3

[23] Jie Lei, Licheng Yu, Mohit Bansal, and Tamara L Berg. Tvqa: Localized, compositional video question answering. *arXiv preprint arXiv:1809.01696*, 2018. 4

[24] Yiheng Li, Ruibing Hou, Hong Chang, Shiguang Shan, and Xilin Chen. Unipose: A unified multimodal framework for human pose comprehension, generation and editing. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 27805–27815, 2025. 2

[25] Bin Lin, Yang Ye, Bin Zhu, Jiaxi Cui, Munan Ning, Peng Jin, and Li Yuan. Video-llava: Learning united visual representation by alignment before projection. *arXiv preprint arXiv:2311.10122*, 2023. 2

[26] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Advances in Neural Information Processing Systems*, pages 34892–34916. Curran Associates, Inc., 2023. 2

[27] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023. 4

[28] Song Liu. Tennis players' hitting action recognition method based on multimodal data. *International Journal of Biometrics*, 16(3-4):317–336, 2024. 1

[29] Muhammad Maaz, Hanoona Rasheed, Salman Khan, and Fahad Shahbaz Khan. Video-chatgpt: Towards detailed video understanding via large vision and language models. *arXiv preprint arXiv:2306.05424*, 2023. 4

[30] Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 2630–2640, 2019. 4

[31] Paritosh Parmar and Brendan Tran Morris. Learning to score olympic events. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 20–28, 2017. 3

[32] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keming Lu, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report, 2025. 5

[33] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in neural information processing systems*, 35:25278–25294, 2022. 4

[34] Zhenhua Tang, Zhaofan Qiu, Yanbin Hao, Richang Hong, and Ting Yao. 3d human pose estimation with spatio-temporal criss-cross attention. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4790–4799, 2023. 2, 5

[35] Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. Movieqa: Understanding stories in movies through question-answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4631–4640, 2016. 4

[36] Zhan Tong, Yibing Song, Jue Wang, and Limin Wang. Videomae: Masked autoencoders are data-efficient learners for self-supervised video pre-training. *Advances in neural information processing systems*, 35:10078–10093, 2022. 2

[37] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 2

[38] Du Tran, Lubomir Bourdev, Rob Fergus, Lorenzo Torresani, and Manohar Paluri. Learning spatiotemporal features with 3d convolutional networks. In *Proceedings of the IEEE international conference on computer vision*, pages 4489–4497, 2015. 2

[39] Chun-Yi Wang, Kalin Guanlun Lai, Hsu-Chun Huang, and Wei-Ting Lin. Tennis player actions dataset for human pose estimation. *Data in Brief*, 55:110665, 2024. 4

[40] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Proceedings of the IEEE international conference on computer vision*, pages 3551–3558, 2013. 2

[41] Julong Wei, Shanshuai Yuan, Pengfei Li, Qingda Hu, Zhongxue Gan, and Wenchao Ding. Occllama: An occupancy-language-action generative world model for autonomous driving, 2024. 2

[42] Jinglin Xu, Sibo Yin, Guohao Zhao, Zishuo Wang, and Yuxin Peng. Fineparser: A fine-grained spatio-temporal action parser for human-centric action quality assessment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14628–14637, 2024. 3

[43] An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025. 2

[44] Weichao Yang. Precise recognition and feature depth analysis of tennis training actions based on multimodal data integration and key action classification. *IEEE Access*, 2025. 1

[45] Jiabo Ye, Haiyang Xu, Haowei Liu, Anwen Hu, Ming Yan, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. mplug-owl3: Towards long image-sequence understanding in multi-modal large language models. *arXiv preprint arXiv:2408.04840*, 2024. 2

[46] Yaqi Zhang, Di Huang, Bin Liu, Shixiang Tang, Yan Lu, Lu Chen, Lei Bai, Qi Chu, Nenghai Yu, and Wanli Ouyang. Motiongpt: Finetuned llms are general-purpose motion generators. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 7368–7376, 2024. 2