electronics

MDPI

Article

BoxingPro: An IoT-LLM Framework for Automated Boxing Coaching via Wearable Sensor Data Fusion

Man Zhu [1], Pengfei Huang [2,*], Xiaolong Xu [1], Houpeng He [1] and Lijie Zhang [1]

[1] School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China; mzhu@njupt.edu.cn (M.Z.)
[2] College of Integrated Circuits, Nanjing University of Aeronautics and Astronautics, Nanjing 211106, China
* Correspondence: pfhuang@nuaa.edu.cn

Abstract

The convergence of Internet of Things (IoT) and Artificial Intelligence (AI) has enabled personalized sports coaching, yet a significant gap remains: translating low-level sensor data into high-level, contextualized feedback. Large Language Models (LLMs) excel at reasoning and instruction but lack a native understanding of physical kinematics. This paper introduces BoxingPro, a novel framework that bridges this semantic gap by fusing wearable sensor data with LLMs for automated boxing coaching. Our core contribution is a dedicated translation methodology that converts multi-modal time-series data (IMU) and visual data (video) into structured linguistic prompts, enabling off-the-shelf LLMs to perform sophisticated biomechanical reasoning without extensive retraining. Our evaluation with professional boxers showed that the generated feedback achieved an average expert rating of over 4.0/5.0 on key criteria like biomechanical correctness and actionability. This work establishes a new paradigm for integrating sensor-based systems with LLMs, with potential applications extending far beyond boxing to any domain requiring physical skill assessment.

Keywords: IMU; pose correction; fitness training; large language models; boxing tutoring

check for updates

Academic Editor: Antoni Morell

Received: 24 September 2025
Revised: 20 October 2025
Accepted: 21 October 2025
Published: 23 October 2025

Citation: Zhu, M.; Huang, P.; Xu, X.; He, H.; Zhang, L. BoxingPro: An IoT-LLM Framework for Automated Boxing Coaching via Wearable Sensor Data Fusion. *Electronics* **2025**, *14*, 4155. https://doi.org/10.3390/electronics14214155

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (https://creativecommons.org/licenses/by/4.0/).

1. Introduction

Boxing has garnered widespread popularity as a combat sport, renowned for delivering an array of physical, psychological, and emotional benefits to its practitioners [1]. In general, boxing boasts a fan base that exceeds half a billion enthusiasts [2]. In conjunction with advancements in exercise equipment and digital training programs, there is a pronounced trend towards at-home exercise routines, with a heightened focus on health and wellness. Yet, this transition is not without its challenges. The absence of on-site coaching means that boxers may struggle to receive the crucial, real-time feedback necessary for refining their technique and form. This deficiency could inadvertently lead to the development of detrimental habits or, worse, increase the risk of injury [3].

The integration of AI technology in sports training has been a rapidly evolving field, with the Internet of Things (IoT) technology emerging as a transformative force. The convergence of IoT and sports training has unlocked the potential for advanced personalized coaching systems [4–6]. These systems not only enable pose recognition but also provide automatic correction feedback, feature of particular importance in the context of boxing training. The past COVID-19 pandemic has necessitated a shift towards individualized

training regimens, making the role of technology in sports coaching more critical than ever [7].

Most research in automatic training feedback in sports has traditionally focused on identifying deviations from standard movements using image processing and deep learning techniques [8–10]. Yet, these approaches have not fully exploited the rich potential of a variety of data sources. The introduction of Generative Pre-trained Transformer 4 (GPT-4) represents a pivotal advancement in artificial intelligence [11], especially in its capacity for broad learning and engaging in chain-of-thought prompting [12]. Integrating GPT-4 with IoT technology holds the promise of significantly enhancing reasoning capabilities in IoT-driven applications [13]. This advancement enables the automatic generation of accurate pose correction recommendations, which is particularly relevant in the context of sports training, including personalized boxing tutoring. This paper aims to investigate how the fusion of GPT-4 and IoT can reshape the field of sports coaching, with a dedicated focus on boxing. This synergy promises to introduce a new paradigm in training methodologies, enhancing the precision and effectiveness of coaching interventions.

In this study, we introduce BoxingPro, an autonomous boxing tutoring system. This system records inertial data from wearable sensors, along with video footage, during the execution of boxing punches by the user. Following a preliminary data preprocessing stage, we conduct analyses of both sensor and video data, and translate the findings into structured language model (LLM) prompts. These prompts are designed to provide personalized feedback and coaching to boxers in a sequential, step-by-step manner. As depicted in Figure 1, the BoxingPro system utilizes two inertial sensors to acquire movement data during boxing exercises. These sensors provide valuable insights into an individual's technique and performance metrics, enabling the system to identify strengths and areas for improvement. By integrating this data with LLMs, the system generates personalized feedback and coaching instructions, tailored to each boxer's unique needs. This system represents a significant leap forward in sports training technology, offering a platform for real-time, data-driven feedback to athletes.

The primary challenge this research addresses is not merely data collection, but the semantic translation of quantitative sensor readings into qualitative, expert-like coaching instructions. This requires a novel pipeline to distill raw kinematics into a form digestible by LLMs. The contributions of this paper are 4-fold:

1. A Novel Translation Methodology for Physical AI: We introduce a structured framework for converting low-level, multi-modal sensor data (IMU kinematics and video-based pose estimation) into high-level, textual representations ("kinematic prompts") that are optimized for LLM comprehension and reasoning. This methodology is the core innovation that enables the application of general-purpose LLMs to specialized physical domains.
2. The BoxingPro System: We instantiate this methodology into a functional system for boxing tutoring, demonstrating its efficacy in providing personalized, step-by-step feedback on technique and form, effectively acting as an automated, data-driven coach.
3. Empirical Validation: We demonstrate the system's performance through rigorous evaluation, including the accuracy of our kinematic feature extraction and, most importantly, the quality of AI-generated feedback as assessed by professional boxers. Our results show that LLMs, when provided with well-structured data, can generate feedback that is not only accurate but also actionable and understandable.
4. A Novel Multimodal Dataset: We gather and will release a unique dataset comprising synchronized IMU sensor data and video recordings of various boxing punches across a diverse group of subjects. This dataset provides a valuable resource for

future research at the intersection of sensor fusion, human activity recognition, and AI-driven coaching.

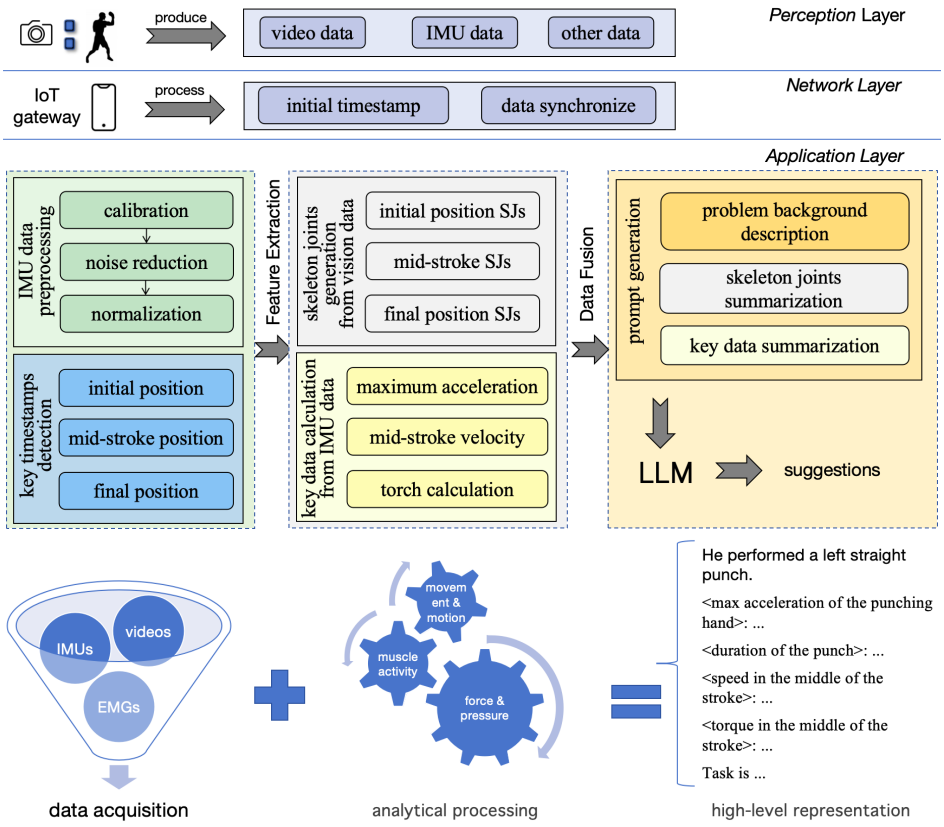


Figure 1. The BoxingPro framework.

2. Related Works

The pursuit of automated, intelligent fitness coaching has evolved through two parallel tracks: vision-based posture analysis and sensor-based activity recognition. The recent emergence of Large Language Models (LLMs) as reasoning engines presents a transformative opportunity to unify these approaches. This section reviews advancements in these domains, highlighting the distinct limitations that our work aims to address. We first examine sensor-free fitness training systems, then explore wearable sensor-based approaches, and finally, investigate the nascent integration of LLMs for multimodal reasoning in physical domains.

2.1. Vision-Based Fitness Training Systems

A significant branch of research has focused on leveraging computer vision for fitness coaching, eliminating the need for physical sensors. These systems primarily use pose estimation algorithms to extract skeletal keypoints from video feeds for analysis.

AIFit [8] advanced the field by performing 3D human body reconstruction from video, providing quantitative feedback by comparing a user's execution to an expert model. Similarly, Pose Trainer [14] utilized OpenPose [10] to analyze exercise postures through vector geometry of body joints. Moving beyond detection, Pose Tutor [9] introduced an angular likelihood mechanism to identify the most critical misaligned joints and offer prioritized corrective feedback. Commercial applications like Zenia have also demonstrated the viability of this technology for domains like yoga through real-time, voice-based guidance. FixMyPose [15] consisted a multilingual (English/Hindi) dataset for pose correction, featuring paired current and target pose images with natural language correction instructions, and

proposed two tasks, a.k.a. pose-correctional captioning and target-pose retrieval. In [16], T. Rangari et al. proposed a video-based system that used OpenPose for 2D joint estimation and an LSTM classifier to recognize exercises and detect correct/incorrect plank postures. Moreover, recent works also adopted 3D pose estimation and encoding techniques [17]. Despite these improvements, the fundamental challenges in dynamic sports and the lack of precise kinematic data remain unresolved by vision-only systems, underscoring the need for a multi-modal approach.

A key limitation of these vision-only systems is their dependence on camera field-of-view and susceptibility to occlusion. Furthermore, they excel at analyzing body position but often lack the precise kinematic data (e.g., acceleration, velocity) required to fully assess the quality of high-speed, dynamic movements like boxing punches.

2.2. Wearable Sensor-Based Activity Recognition

To overcome the limitations of vision, a complementary body of work employs wearable sensors to capture rich kinematic data directly from the body. These systems are often more robust to environmental constraints like lighting and occlusion.

In yoga, YogaHelp [4] used motion sensors and a deep learning model to recognize and evaluate the correctness of a multi-step yoga sequence. For general gym activities, GPARMF [5] proposed a hybrid framework that fuses sensor data for accurate activity recognition. Other approaches, like that of Gochoo et al. [6], have focused on privacy-preserving sensing using low-resolution infrared sensors paired with deep neural networks for posture recognition. The most relevant to our work are efforts that explicitly fuse sensor modalities. FusePose [18], for instance, integrates IMU and vision data to enhance the accuracy of pose estimation itself. M. Wozniak [19] presented a mixed reality system that uses a recurrent neural network and two body sensors to detect unusual poses, providing ad-hoc safety monitoring for workers and elderly people in various environments without needing special infrastructure. In [20], Phukan et al. presented an optimized convolutional neural network (CNN)-based human activity recognition method using accelerometer data. Previous works also used deep learning models for time-series analysis from wearable sensors, with convolutional and recurrent neural networks for activity recognition [21].

Directly in boxing, Boxersense [22] demonstrated the feasibility of classifying punch types using acceleration and gyroscope data from a single smartwatch. However, a common shortcoming of these sensor-based systems is their primary focus on activity recognition and classification. They effectively answer "what" activity was performed but often lack the framework to generate nuanced, contextual coaching feedback to explain "how" to improve, limiting their utility as personalized tutors.

Wearable sensor-based approaches are suitable for analyzing the dynamics of boxing. However, the focus of this line of research has predominantly been on activity recognition and classification—accurately identifying what action was performed [22]. While this is a valuable first step, it falls short of personalized coaching. This gap between low-level data recognition and high-level instructional feedback represents a significant limitation in existing sensor-based tutors.

2.3. Towards Integrated Coaching: The Role of Large Language Models

The application of Large Language Models (LLMs) extends far beyond text, showing immense promise for reasoning about the physical world through sensor data. Recent research has begun to explore LLMs as cognitive engines for the Internet of Things (IoT).

Studies like IoT-LLM [23] have focused on augmenting LLMs with task reasoning capabilities by embedding sensor data into their context. An et al. [13] powerfully framed IoT sensors as the sensory organs providing perception for LLMs, which act as the central

brain for interpretation and decision-making. Similarly, Xu et al. [24] explored LLMs for penetrative reasoning in IoT environments.

However, the application of this IoT-LLM paradigm to complex kinematic analysis and personalized feedback generation in sports science remains largely unexplored. The challenge of grounding LLMs in physical reality is an active research area, often explored in domains like embodied task planning [25]. Existing works have not yet fully leveraged the unique capabilities of LLMs to synthesize multi-modal data—fusing quantitative kinematic metrics from sensors with qualitative postural analysis from vision—to generate holistic, natural language coaching instructions. This gap presents a significant opportunity, which our work seizes by proposing a novel pipeline that translates raw sensor and video data into high-level textual feedback for boxing, effectively creating a personalized AI coach.

3. A Novel Translation Methodology for Physical AI

Figure 1 illustrates the proposed pipeline for translating low-level, multi-modal sensor data into high-level, actionable textual feedback. This methodology is core to our concept of "Physical AI", where intelligent interpretation of physical actions bridges the gap between raw data and human-understandable coaching. This pipeline is based on a classic IoT stack, designed to bridge the physical world of boxing kinematics with the digital realm of AI-driven coaching. The architecture consists of three layers: (1) Perception Layer: This layer comprises the physical sensors that perceive the athlete's actions. It includes wearable IMU sensors, and vision sensors. (2) Network Layer: This layer handles data communication and transmission. The IMUs stream data via Bluetooth to a central smartphone, which acts as an IoT gateway. The gateway synchronizes the IMU data with the video feed, applies initial timestamps, and transmits the aggregated, multi-modal data stream to the server via a WIFI connection. (3) Application Layer: Residing on the cloud server, this application layer executes the comprehensive analytical processing pipeline. The pipeline begins with multi-modal feature extraction—from IMU data, we derive kinematic metrics including acceleration magnitude, angular velocity, and motion trajectory; from video data, we extract skeletal keypoints and biomechanical features using pose estimation. This is followed by temporal-spatial data fusion that synchronizes and integrates these heterogeneous data streams into unified kinematic representations. The layer then interfaces with the Large Language Model, transforming these structured physical analytics into contextualized coaching instructions through carefully designed prompt templates.

The translation pipeline is structured into three integral modules: data acquisition, analytical processing, and linguistic representation generation.

Perception layer data acquisition. The initial phase involves multi-modal data collection from participants equipped with wireless Inertial Measurement Units (IMUs) affixed to both hands. These IMUs stream data via Bluetooth to a mobile device running the custom BoxingPro application. The application simultaneously records high-frame-rate (200 FPS) video of the boxer's performance, to adequately capture the high-speed kinematics of boxing punches, ensuring temporal synchronization between the kinematic and visual data streams. The camera was positioned from the side view at chest height in order to optimally capture the kinematic chain from rear foot to lead hand during punching motions. The IMU data comprises nine distinct kinematic attributes: tri-axial acceleration (accX, accY, accZ), tri-axial angular velocity (gyrX, gyrY, gyrZ), and orientation angles (pitchAngle, rollAngle, rawAngle). A synchronized timestamp for each data point is critical for subsequent processing. This aggregated dataset is then transmitted to a central server for robust analysis.

Application layer analytical processing. Upon receipt, the server processes the synchronized IMU and video data through a sequence of steps to extract semantically meaningful features.

1. IMU data preprocessing. The raw IMU signals are refined to ensure data integrity and uniformity. This involves noise reduction, sensor bias correction, and synchronization. We apply a 5-order Butterworth low-pass filter [26] to each kinematic attribute to eliminate high-frequency noise while preserving signal integrity, followed by standardization. The resulting preprocessed dataset is denoted as D.
2. Key timestamps detection. Critical events within a punch cycle are identified by detecting key timestamps. This process begins by computing the first discrete difference of the preprocessed dataset D (excluding the timestamp column), yielding a differential dataset D^d that highlights changes in kinematic states. Minor fluctuations are filtered out by removing entries without values fall above 0.001, producing D'. The key timestamps—indicating the initial position, mid-stroke (impact point), and final position—are then algorithmically identified from D' as outlined in Table 1. The mid-stroke is pinpointed at the timestamp of maximum resultant acceleration. The initial and final positions are defined as the timestamps of minimum acceleration immediately preceding and succeeding the mid-stroke, respectively.
3. Skeleton joints generation. For the video frames corresponding to the key timestamps, we perform human pose estimation to generate a skeletal model of the boxer. This is achieved using MoveNet [27], which was selected for its effective balance of speed and accuracy. MoveNet is a lightweight convolutional neural network built upon a MobileNetV2 backbone. The model accepts 192 × 192 × 3 RGB images as input and produces 17 human body keypoints, each accompanied by coordinate confidence scores. Key technical specifications include: 3.4 million parameters, an inference speed of approximately 30 FPS on mobile devices, and a mAP of 73.8% on the COCO validation dataset. The extracted skeletal joints, as visualized in Figure 2, provide crucial contextual information on body posture and biomechanics that is absent from IMU data alone.
4. Biomechanical metric calculation. Utilizing the preprocessed data D and the key timestamps, we calculate quantitative metrics that define punch quality (Table 2). These include maximum acceleration (indicative of explosive power), punch speed (calculated by integrating acceleration over time from initial position until the mid-stroke), angular momentum ($L = W \cdot D_{gyr}$), and derived torque ($\tau = \frac{dL}{dt}$), which is calculated as it is a known determinant of punching force generation [28]. These metrics offer a objective foundation for assessing technique and generating targeted feedback.

Table 1. Key timestamps calculation.

Key Timestamp	Equation
mid stroke	$ts[max(\lvert D^d_{acc}\rvert)]$
initial position	$ts[min(\lvert D^d_{acc}[: mid - ts]\rvert)]$
final position	$ts[min(\lvert D^d_{acc}[mid - ts :]\rvert)]$

Table 2. Key data calculation. T^d is the discrete difference of the timestamp column in D, gyr is short for gyroscope, and L^d is the discrete difference of L.

Key Data	Equation
max accelerations	$[max(\lvert D^d_{accX}\rvert), max(\lvert D^d_{accY}\rvert), max(\lvert D^d_{accZ}\rvert)]$
max speed	$D^d_{acc}[: midstroke] \cdot T^d$
angular momentum	$L = W \cdot D_{gyr}$
torque	$\frac{dL}{dT}$

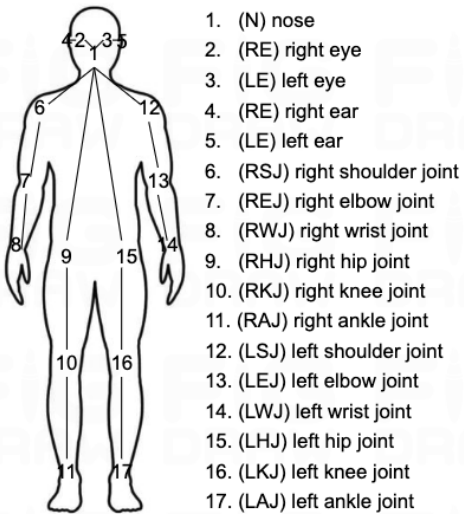


Figure 2. Body skeleton joints.

Linguistic Representation Generation. The final module translates the analytical features into natural language. Structured prompts are generated to serve as personalized coaching instructions. Each prompt is crafted to: (a) contextualize the performance within the specific punch type, (b) summarize the key biomechanical findings from the analysis, and (c) provide a concrete, actionable recommendation for improvement. An example of this structured feedback for a left uppercut is depicted in Figure 3. This step effectively closes the loop, transforming quantitative data into qualitative, human-centric coaching guidance.

problem background description	A boxer executed a LU. Through the analysis based on the datasets collected from the IMUs held in his left and right hands, we find the following characteristics:
key data summarization	<max acceleration of the punching hand>: X-axis 3.33900088697305 m/s2, Y-axis 1.435110549170777 m/s2, Z-axis 7.825897907027875 m/s2 <duration of the punch>: 910.0 milliseconds <speed in the middle of the stroke>: X-axis 0.49617054809713307 m/s, Y-axis 2.096367590069099 m/s, Z-axis 3.1683122035207405 m/s <torque in the middle of the stroke>: X-axis -0.03608097768703781 Nm, Y-axis 0.06384601775624332 Nm, Z-axis 0.07344781064518338Nm
task description	Your task is: Please analyze this punch according to the characteristics, and give specific advices of improvement is possible.

Figure 3. Linguistic representation representation. LU is the acronym of Lead Uppercut. All the acronyms were denoted before all tasks.

4. Evaluation

This section presents a comprehensive evaluation of the BoxingPro system, designed to quantitatively and qualitatively assess its core functionalities. We measure the accuracy of our low-level kinematic processing pipeline and evaluate the quality of the high-level, personalized feedback generated by the integrated Large Language Models (LLMs). The evaluation is structured around three primary research questions (RQ):

- RQ1: Temporal Accuracy: How precisely does the system detect key biomechanical events (initial, mid-stroke, final positions) in a punch sequence?

- RQ2: Feedback Quality: How effective, accurate, and actionable is the coaching feedback generated by different LLMs?
- RQ3: The Impact of Fine-Tuning: Does domain-specific fine-tuning of an LLM improve the quality of generated boxing feedback?

These structured prompts are then fed into a Large Language Model (e.g., GPT-4) to generate personalized coaching advice. The quality of this AI-generated feedback was quantitatively evaluated by professional boxers across five defined criteria: Biomechanical Correctness, Quantifiable Metrics, Operability, Feedback Timing, and Understandability, as detailed in Section "Performance Metrics". The LLM synthesizes the structured data points into coherent paragraphs, mimicking the explanatory style of a human coach.

4.1. Experimental Setup

4.1.1. Dataset and Participants

To ensure a robust and generalizable evaluation, we collected a multimodal dataset comprising synchronized video and IMU data. The IMU is a Wit-Motion sensor (model JY61). It streams 9-axis kinematic data—including tri-axial acceleration ($\pm$16 g), angular velocity ($\pm 2000^\circ/s$), and orientation angles—at a configurable sampling rate up to 200 Hz via Bluetooth transmission.

We recruited 18 volunteers (4 female, 14 male) aged 5 to 40. Participants were classified into three distinct categories based on formal training experience: (1) Novices (N, n = 13): Individuals with no formal boxing training (0 months of structured coaching). (2) Beginner (B, n = 1): A participant with 3–6 months of formal boxing training under certified coaching, demonstrating basic technique understanding but limited practical application. (3) Professionals (P, n = 4): Competitive boxers with minimum 3 years of continuous training and active competition experience in regional or national tournaments. Demographic details are provided in Table 3. Each participant performed six fundamental punches (Jab, Cross, Lead/Rear Hook, Lead/Rear Uppercut) three times each, resulting in a total of 324 unique punch instances with paired data. This diversity in age, skill, and physique helps validate the system's robustness across a varied population.

Table 3. Statistics of volunteers. N is short for new, B is short for beginner, and P is short for professional.

ID	Sex	Age	Height [m]	Weight [kg]	BMI [kg/m^2]	Level
1	M	19	1.72	56	18.9	N
2	M	20	1.75	64	20.9	N
3	M	20	1.80	68	21.1	N
4	M	21	1.70	47	16.3	N
5	M	20	1.68	49	17.4	N
6	F	20	1.51	40	17.4	N
7	F	19	1.67	53	19.0	N
8	M	18	1.75	70	22.9	P
9	M	18	1.80	75	23.1	P
10	M	18	1.75	72	23.5	P
11	M	18	1.75	68	22.2	P
12	F	39	1.69	64	22.41	N
13	M	40	1.75	80	26.1	N
14	M	7	1.35	30	16.5	B
15	M	5	1.15	20	15.1	N
16	M	18	1.75	65	21.2	N
17	F	18	1.75	60	19.6	N
18	F	18	1.75	58	18.9	N

4.1.2. Models and Comparison

To evaluate feedback generation, we compared multiple state-of-the-art LLMs:

- Llama2-7B [29]: A transformer-based model with 7 billion parameters, 32 layers, 4096 hidden dimensions, and 32 attention heads, using SwiGLU activation function and pre-normalization configuration.
- ChatGLM [30]: A bilingual GLM architecture with 6 billion parameters, 28 transformer layers, and 4096 hidden dimensions, optimized for dialogue tasks.
- DeepSeek-V3 [31]: A mixture-of-experts model with 67 billion total parameters (37 billion active), 54 layers, 7680 hidden dimensions, and 60 attention heads, supporting 128 K context length.
- Llama2-7B (Fine-Tuned): We fine-tuned the base model using Low-Rank Adaptation (LoRA) [32] with rank $r = 16$, trained for 1000 steps with learning rate 3×10^{-4} on 1500 boxing-specific prompt-suggestion pairs.

All models were deployed on an Alibaba cloud server in Hangzhou, China, with 8 vCPUs, 64 GB RAM and 2 NVIDIA P100 GPUs, running Ubuntu 18.04 and served via the vLLM engine for maximum throughput.

The prompts consisted of the structured data (kinematic metrics, skeletal angles) output by our pipeline, and the suggestions were crafted and validated by professional boxers.

4.2. Boxing Poses

The selection of boxing poses for this study was deliberate and comprehensive, encompassing a range of defensive maneuvers and offensive strikes. The poses include the following:

- Jab: A straight punch delivered with the lead hand.
- Cross: A straight punch delivered with the rear hand.
- Lead Hook: A circular punch thrown with the lead hand.
- Rear Hook: A circular punch thrown with the rear hand.
- Lead Uppercut: An upward punch delivered with the lead hand.
- Rear Uppercut: An upward punch delivered with the rear hand.

The dynamic and transient nature of these poses presents a unique challenge for chracterization and correction. Figure 4 illustrates the iconic moments of each boxing pose as executed by a professional boxer, capturing the essence of each movement.



Figure 4. Iconic moments of 6 boxing poses executed by Liang Tang, a professional Chinese boxer.

Performance Metrics

Evaluation was conducted along two distinct dimensions:

Temporal Detection Accuracy (RQ1): The accuracy of detecting key timestamps (mid-stroke, initial, final positions) was assessed using the Mean Squared Error (MSE) between the system-detected time (t_{id}^c) and the manually annotated ground-truth time (t_{ia}^c) for each data instance i and category c:

$$\text{MSE} = \frac{1}{|D|} \sum_{i \in D} (t_{id}^c - t_{ia}^c)^2 \quad (1)$$

where D is the entire dataset. Lower MSE indicates higher precision.

Feedback Quality Assessment (RQ2, RQ3): The quality of the LLM-generated feedback was evaluated by our panel of four professional boxers (IDs 8-11 in Table 3). They scored each suggestion on a Likert scale from 0 (Very Poor) to 5 (Excellent) across five criteria:

- Biomechanical Correctness (B): Technical accuracy of the advice.
- Quantifiable Metrics (Q): Use of data-driven, measurable insights.
- Operability (O): Actionability and ease of implementing the suggestion.
- Feedback Timing (F): Perceived relevance of the feedback to the phase of the punch.
- Understandability (U): Clarity and conciseness of the language.

4.3. Results and Analysis

4.3.1. Key Timestamps Detection (RQ1)

The results for detecting the critical *mid-stroke* timestamp are summarized in Table 4 and visualized in Figure 5. Our analysis reveals a clear performance asymmetry between dominant and non-dominant hands.

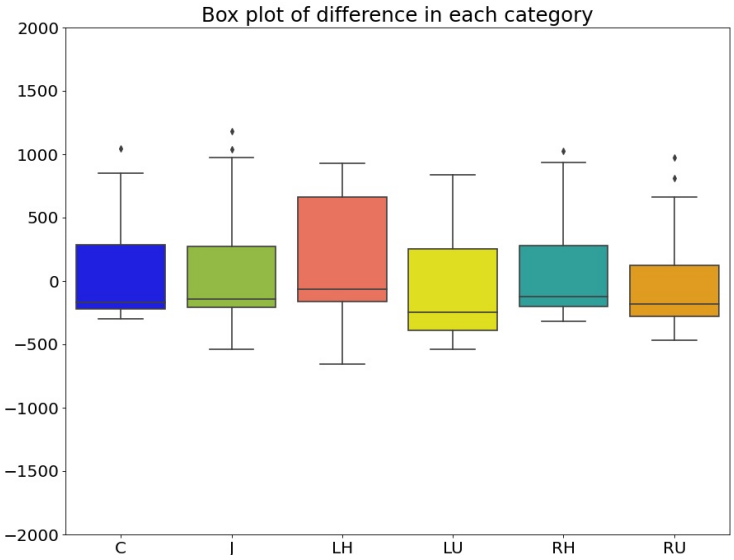


Figure 5. Box plot of annotation vs. detection timestamp differences (milliseconds) by category.

The Right Uppercut (RU) and Cross (C) demonstrated the highest detection accuracy (lowest MSE), attributed to the more pronounced and consistent kinematic signatures (e.g., sharper acceleration peaks, cleaner rotation) generated by the dominant side in right-handed individuals. Conversely, left-hand strikes (LH, LU) showed higher error, with the Jab (J) exhibiting the highest variability. This is likely due to its linear trajectory and lower peak acceleration compared to the rotational hooks and uppercuts, making its key kinematic events slightly less distinct for the algorithm to pinpoint.

These results confirm that the pipeline effectively captures major biomechanical events, with performance influenced by punch kinematics and user handedness.

Table 4. Mean Squared Error (MSE) and Standard Deviation (SD) for Mid-Stroke Timestamp Detection across Punch Types. (J: Jab, C: Cross, LH: Left Hook, RH: Right Hook, LU: Left Uppercut, RU: Right Uppercut).

Punch Type	Average MSE (ms)	Standard Deviation
J	0.218	0.351
C	0.151	0.233
LH	0.222	0.297
RH	0.162	0.271
LU	0.196	0.196
RU	0.134	0.182

4.3.2. Feedback Efficacy (RQ2 & RQ3)

The results of the expert evaluation are presented in Table 5. Two key findings emerge:

Superiority of Advanced and Fine-Tuned Models: The general-purpose DeepSeek-V3 (DS-V3) and the fine-tuned Llama2-7B (FT) model significantly outperformed the base Llama2-7B and ChatGLM across all five dimensions. This is particularly evident in Quantifiable Metrics (Q) and Understandability (U), where both scored above 4.0. The fine-tuned model achieved the highest scores in Operability (O) (4.3) and Understandability (U) (4.5), demonstrating that domain-specific adaptation makes feedback more actionable and clear.

Identification of a Key Limitation: The dimension of Feedback Timing (F) received the lowest scores across all models, including the best-performing ones (DS-V3 = 4.0, Llama2-7B FT = 3.5). As further illustrated in Figure 6, this was consistent across all punch types. This reflects the inherent computational latency of using LLMs for real-time feedback generation, a critical challenge for in-session coaching.

Figure 7 provides a qualitative comparison, showing how the fine-tuned model generates more specific, technically accurate, and structured feedback compared to a generic model like ChatGLM. In the figure, "###" denotes a new section in the outputs, a pair of "**" embrace an item title, and a single "**" starts a sub-item.

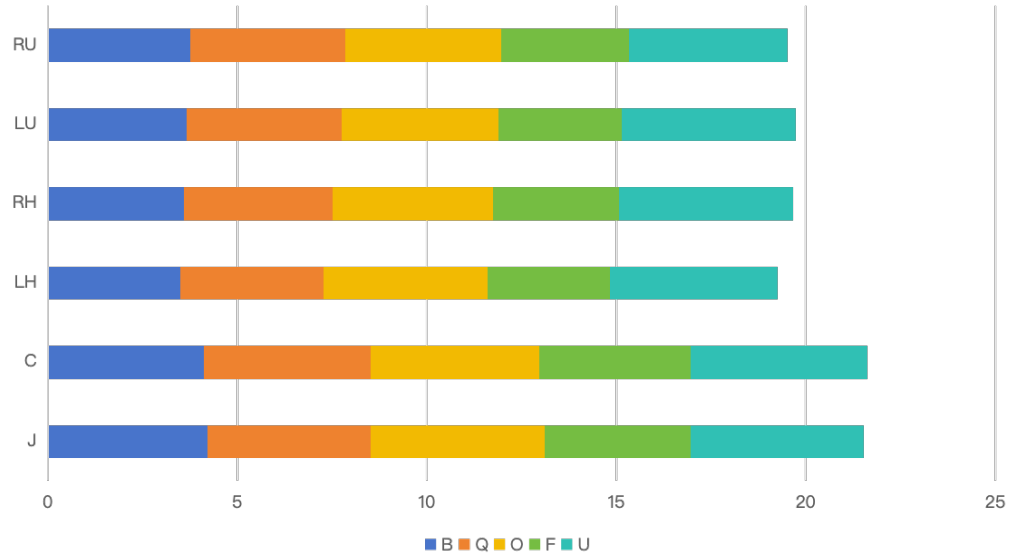


Figure 6. Performance evaluation of fine-tuned Llama2-7B: aggregate scores across 5 feedback categories for different punch types.

```
### Specific Advice for Improvement:

1. **Enhance Vertical Power:**
   - **Focus on Leg Drive: ...
   - **Increase Y-axis Acceleration:...

2. **Optimize Forward Movement:**
   - **Improve Hip Rotation:...
   - **Stance Adjustment:...

3. **Balance Speed and Power:**
   - **Slight Increase in Duration:...
   - **Plyometric Training:...

4. **Strengthen Core and Upper Body:**
   - **Core Stability Exercises:...
   - **Upper Body Strength Training:...

### Conclusion:
The boxer demonstrates a decent execution of the Lead Uppercut with good speed and moderate power.
However, by focusing on enhancing vertical power, optimizing forward movement, refining technique, balancing
speed and power, and strengthening the core and upper body, the boxer can significantly improve the
effectiveness of the punch. Regular monitoring and adjustments based on continuous data analysis will be crucial
for sustained improvement.
```

```
What the data shows:
- Duration: Around 0.86 seconds; decent but could be faster.

Why it matters:
- Speed improvements make your punch more effective, especially in close exchanges.
- Increasing horizontal acceleration and reducing punch time will make your hook more explosive.

What to do:
- Focus on quick hip rotation and sharp shoulder turn.
- Use explosive drills like medicine ball throws or shadowboxing with a timer.
- Keep your stance stable and pivot sharply on your lead foot to add power.

Expected results:
- Cutting your punch duration from about 0.86 seconds to under 0.5 seconds will result in more impactful, quicker
hooks that are difficult to read and block.
```

Figure 7. Example suggestions using ChatGLM and fine-tuned Llama2-7B respectively.

Table 5. Expert Evaluation Scores (0–5) of Feedback Generated by Different LLMs. (B: Biomechanical Correctness, Q: Quantifiable Metrics, O: Operability, F: Feedback Timing, U: Understandability).

Dimension	Llama2 7B	ChatGLM	DS-V3	Llama2 7B FT
B	3.2	3.5	4.1	3.8
Q	2.9	3.1	4.2	4.1
O	3.0	3.4	3.9	4.3
F	3.1	3.0	4.0	3.5
U	3.3	3.6	4.3	4.5

4.4. Discussion and Limitations

The evaluation confirms the core viability of the BoxingPro pipeline. The system accurately detects key kinematic events and, when paired with a powerful or fine-tuned LLM, generates high-quality, expert-validated feedback.

However, the study highlights two primary limitations:

Real-Time Performance: The latency in feedback generation, as captured by the low feedback timing scores, is the most significant barrier to deployment for real-time coaching. Future work must explore model optimization, quantization, and potentially different architectural choices to mitigate this.

Data Scale and Bias: While the dataset is diverse, its size (324 instances) is limited. A larger dataset would improve the robustness of the kinematic analysis and the quality of the fine-tuning process for the LLM. Besides, the dataset is potential for bias by the gender and the age of the participants.

Despite these limitations, the results are highly promising. The system demonstrates a successful proof-of-concept for fusing low-level sensor data with high-level AI reasoning to create a personalized boxing tutor.

5. Conclusions

This paper presented BoxingPro, a novel framework that fuses data from wearable IoT sensors with the reasoning capabilities of large language models (LLMs) to create an automated boxing coach. Our central contribution is a dedicated translation methodology that converts multi-modal, time-series kinematic data and video-based pose estimates into structured textual prompts. This pipeline effectively bridges the critical semantic gap between quantitative sensor readings and qualitative, human-understandable coaching instructions, establishing a new paradigm for 'Physical AI' tutors.

The evaluation demonstrated the pipeline's effectiveness in two perspectives: Firstly, it achieved high temporal accuracy in detecting key biomechanical events (e.g., mid-stroke impact); secondly, and more importantly, the feedback generated by LLMs like DeepSeek-V3 and a fine-tuned Llama2-7B received high scores from professional boxers across criteria such as biomechanical correctness, actionability, and understandability. This validates the core premise that LLMs can function as powerful reasoning engines for physical skill assessment when provided with appropriately structured data.

In the future, we will focus on: (1) optimizing the LLM component for real-time latency, (2) expanding the dataset to include more diverse boxers and punch variations, and (3) exploring the integration of this system into augmented reality (AR) glasses for immersive in-training feedback.

The BoxingPro framework not only advances sports technology but also provides a scalable blueprint for integrating sensor-driven systems with LLMs across various domains requiring physical skill assessment.

Author Contributions: Conceptualization, M.Z.; Methodology, M.Z. and X.X.; Software, H.H. and L.Z.; Validation, L.Z.; Investigation, P.H.; Resources, M.Z. and H.H.; Data curation, L.Z.; Writing—original draft, M.Z.; Writing—review & editing, P.H.; Visualization, P.H.; Supervision, P.H.; Funding acquisition, M.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported by the National Natural Science Foundation of China under Grant No. U23A20307.

Data Availability Statement: The data presented in this study are available on request from the corresponding author.

Acknowledgments: We thank Liang Tang and Chao Tang of the Chang Sheng Boxing Club for facilitating the data collection. We also extend our gratitude to professional boxer Marouane Harrach for his crucial role in evaluating the LLM-generated feedback.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Worsey, M.T.; Espinosa, H.G.; Shepherd, J.B.; Thiel, D.V. An evaluation of wearable inertial sensor configuration and supervised machine learning models for automatic punch classification in boxing. *IoT* **2020**, *1*, 360–381. [CrossRef]
2. Statista. U.S. Americans Who Did Boxing. 2023. Available online: https://www.statista.com/statistics/191905/participants-in-boxing-in-the-us-since-2006/ (accessed on 28 December 2024).
3. Alevras, A.J.; Fuller, J.T.; Mitchell, R.; Lystad, R.P. Epidemiology of injuries in amateur boxing: A systematic review and meta-analysis. *J. Sci. Med. Sport* **2022**, *25*, 995–1001. [CrossRef]
4. Gupta, A.; Gupta, H.P. YogaHelp: Leveraging Motion Sensors for Learning Correct Execution of Yoga With Feedback. *IEEE Trans. Artif. Intell.* **2021**, *2*, 362–371. [CrossRef]
5. Qi, J.; Yang, P.; Hanneghan, M.; Tang, S.; Zhou, B. A hybrid hierarchical framework for gym physical activity recognition and measurement using wearable sensors. *IEEE Internet Things J.* **2018**, *6*, 1384–1393. [CrossRef]
6. Gochoo, M.; Tan, T.H.; Huang, S.C.; Batjargal, T.; Hsieh, J.W.; Alnajjar, F.S.; Chen, Y.F. Novel IoT-based privacy-preserving yoga posture recognition system using low-resolution infrared sensors and deep learning. *IEEE Internet Things J.* **2019**, *6*, 7192–7200. [CrossRef]

7. Wang, J.; Qiu, K.; Peng, H.; Fu, J.; Zhu, J. Ai coach: Deep human pose estimation and analysis for personalized athletic training assistance. In Proceedings of the 27th ACM International Conference on Multimedia, Nice, France, 21–25 October 2019; pp. 374–382.
8. Fieraru, M.; Zanfir, M.; Pirlea, S.C.; Olaru, V.; Sminchisescu, C. Aifit: Automatic 3d human-interpretable feedback models for fitness training. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, Virtual, 19–25 June 2021; pp. 9919–9928.
9. Dittakavi, B.; Bavikadi, D.; Desai, S.V.; Chakraborty, S.; Reddy, N.; Balasubramanian, V.N.; Callepalli, B.; Sharma, A. Pose tutor: An explainable system for pose correction in the wild. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, New Orleans, LA, USA, 18–24 June 2022; pp. 3540–3549.
10. Cao, Z.; Simon, T.; Wei, S.E.; Sheikh, Y. Realtime multi-person 2d pose estimation using part affinity fields. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 21–26 July 2017; pp. 7291–7299.
11. Zhao, W.X.; Zhou, K.; Li, J.; Tang, T.; Wang, X.; Hou, Y.; Min, Y.; Zhang, B.; Zhang, J.; Dong, Z.; et al. A survey of large language models. *arXiv* **2023**, arXiv:2303.18223. [PubMed]
12. Feng, G.; Zhang, B.; Gu, Y.; Ye, H.; He, D.; Wang, L. Towards revealing the mystery behind chain of thought: A theoretical perspective. In *Advances in Neural Information Processing Systems, Proceedings of the Conference and Workshop on Neural Information Processing Systems 2024, Vancouver, BC, Canada, 10–15 November 2024*; NeurIPS Foundation: La Jolla, CA, USA, 2024; Volume 36.
13. An, T.; Zhou, Y.; Zou, H.; Yang, J. IoT-LLM: Enhancing Real-World IoT Task Reasoning with Large Language Models. *arXiv* **2024**, arXiv:2410.02429.
14. Chen, S.; Yang, R.R. Pose trainer: Correcting exercise posture using pose estimation. *arXiv* **2020**, arXiv:2006.11718. [CrossRef]
15. Kim, H.; Zala, A.; Burri, G.; Bansal, M. FIXMYPOSE: Pose Correctional Captioning and Retrieval. In Proceedings of the Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual, 2–9 February 2021; AAAI Press: Washington, DC, USA, 2021; pp. 13161–13170. [CrossRef]
16. Rangari, T.; Kumar, S.; Roy, P.P.; Dogra, D.P.; Kim, B. Video based exercise recognition and correct pose detection. *Multim. Tools Appl.* **2022**, *81*, 30267–30282. [CrossRef]
17. Debnath, B.; O'Brient, M.; Kumar, S.; Behera, A. Attention-driven body pose encoding for human activity recognition. In Proceedings of the 2020 25th International Conference on Pattern Recognition (ICPR), Virtual, 10–15 January 2021; IEEE: New York, NY, USA, 2021; pp. 5897–5904.
18. Bao, Y.; Zhao, X.; Qian, D. FusePose: IMU-Vision Sensor Fusion in Kinematic Space for Parametric Human Pose Estimation. *IEEE Trans. Multimed.* **2023**, *25*, 7736–7746. [CrossRef]
19. Wozniak, M.; Wieczorek, M.; Silka, J.; Polap, D. Body Pose Prediction Based on Motion Sensor Data and Recurrent Neural Network. *IEEE Trans. Ind. Inform.* **2021**, *17*, 2101–2111. [CrossRef]
20. Phukan, N.; Mohine, S.; Mondal, A.; Manikandan, M.S.; Pachori, R.B. Convolutional neural network-based human activity recognition for edge fitness and context-aware health monitoring devices. *IEEE Sens. J.* **2022**, *22*, 21816–21826. [CrossRef]
21. Hammerla, N.Y.; Halloran, S.; Pl′otz, T. Deep, convolutional, and recurrent models for human activity recognition using wearables. *arXiv* **2016**, arXiv:1604.08880. [CrossRef]
22. Hanada, Y.; Hossain, T.; Yokokubo, A.; Lopez, G. Boxersense: Punch detection and classification using IMUs. In *Sensor-and Video-Based Activity and Behavior Computing, Proceedings of 3rd International Conference on Activity and Behavior Computing (ABC 2021), Online, 22–23 October 2021*; Springer: Berlin/Heidelberg, Germany, 2021; pp. 95–114.
23. Ji, S.; Zheng, X.; Wu, C. HARGPT: Are LLMs Zero-Shot Human Activity Recognizers? *arXiv* **2024**, arXiv:2403.02727. [CrossRef]
24. Xu, H.; Han, L.; Yang, Q.; Li, M.; Srivastava, M. Penetrative ai: Making llms comprehend the physical world. In Proceedings of the 25th International Workshop on Mobile Computing Systems and Applications, San Diego, CA, USA, 28–29 February 2024; pp. 1–7.
25. Wu, Z.; Wang, Z.; Xu, X.; Lu, J.; Yan, H. Embodied task planning with large language models. *arXiv* **2023**, arXiv:2307.01848. [CrossRef]
26. Yu, B.; Gabriel, D.; Noble, L.; An, K.N. Estimate of the optimum cutoff frequency for the Butterworth low-pass digital filter. *J. Appl. Biomech.* **1999**, *15*, 318–329. [CrossRef]
27. Bajpai, R.; Joshi, D. Movenet: A deep neural network for joint profile prediction across variable walking speeds and slopes. *IEEE Trans. Instrum. Meas.* **2021**, *70*, 2508511. [CrossRef]
28. Scattone-Silva, R.; Lessi, G.; Lobato, D.; Serrão, F. Acceleration time, peak torque and time to peak torque in elite karate athletes. *Sci. Sport.* **2012**, *27*, e31–e37. [CrossRef]
29. Touvron, H.; Lavril, T.; Izacard, G.; Martinet, X.; Lachaux, M.A.; Lacroix, T.; Rozière, B.; Goyal, N.; Hambro, E.; Azhar, F.; et al. Llama: Open and efficient foundation language models. *arXiv* **2023**, arXiv:2302.13971. [CrossRef]
30. Zeng, A.; Liu, X.; Du, Z.; Wang, Z.; Lai, H.; Ding, M.; Yang, Z.; Xu, Y.; Zheng, W.; Xia, X.; et al. Glm-130b: An open bilingual pre-trained model. DeepSeek-V3 Technical Report. *arXiv* **2022**, arXiv:2210.02414.

31. DeepSeek-AI; Liu, A.; Feng, B.; Xue, B.; Wang, B.; Wu, B.; Lu, C.; Zhao, C.; Deng, C.; Zhang, C.; et al. DeepSeek-V3 Technical Report. *arXiv* **2024**, arXiv:2412.19437.
32. Hu, E.J.; Shen, Y.; Wallis, P.; Allen-Zhu, Z.; Li, Y.; Wang, S.; Wang, L.; Chen, W. Lora: Low-rank adaptation of large language models. In Proceedings of the International Conference on Learning Representations, Virtual, 25–29 April 2022; Volume 1, p. 3.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.