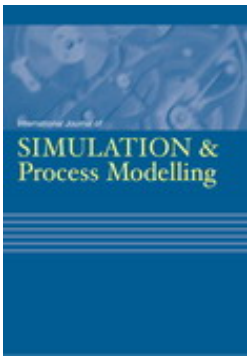


International Journal of Simulation and Process Modelling

ISSN online: 1740-2131 - ISSN print: 1740-2123
https://www.inderscience.com/ijspm

Multimodal pose estimation and simulation modelling for real-time human motion analysis

Dongsheng Chen, Zhen Ni, Wei Huang

DOI: 10.1504/IJSPM.2025.10074125

Article History:

Received:	03 July 2025
Last revised:	11 August 2025
Accepted:	28 August 2025
Published online:	24 October 2025

Multimodal pose estimation and simulation modelling for real-time human motion analysis

Dongsheng Chen

School of Sports and Health,
Guangxi College for Preschool Education,
Guangxi 530022, China
Email: cdsdqm@126.com

Zhen Ni

School of Physical Education and Health,
Nanning Normal University,
Nanning 530001, China
Email: 18778012817@126.com

Wei Huang*

School of Sports and Health,
Guangxi College for Preschool Education,
Guangxi 530022, China
Email: weihuang19810728@163.com
*Corresponding author

Abstract: To ensure safe and effective campus physical activities, this pioneering study proposes an innovative real-time sports pose recognition framework integrated with simulation-oriented process modelling, aligning with the core scope of dynamic motion analysis. The framework features a sophisticated multimodal architecture that fuses visual and inertial data across four interconnected layers, while embedding simulation-driven process modelling to capture the spatiotemporal dynamics of human motion. Enhanced spatiotemporal alignment mechanisms enable precise extraction of key biomechanical features, which are further refined through optimised Relief F algorithm for critical motion feature selection. A particle swarm-optimised graph convolutional network (PSO-AGCN) leverages simulated motion topology variations to process these features efficiently for pose classification. Evaluations on Human3.6M and a college sports dataset show 96.7% accuracy, 42.3% reduced occlusion errors, and 38 FPS operation, highlighting robustness and real-time performance, with simulation enhancing analysis interpretability.

Keywords: multimodal pose estimation; simulation modelling; real-time motion analysis; graph convolutional networks; sports simulation; campus sports analytics.

Reference to this paper should be made as follows: Chen, D., Ni, Z. and Huang, W. (2025) 'Multimodal pose estimation and simulation modelling for real-time human motion analysis', *Int. J. Simulation and Process Modelling*, Vol. 22, No. 5, pp.1–10.

Biographical notes: Dongsheng Chen received her PhD degree from the Jose Rizal University in 2022. Currently, she is working in the Guangxi College for Preschool Education. Her research interests include physical education and training and ed.

Zhen Ni received his Master's degree from the Beijing Sport University in 2006. Since 1999, he works in the Nanning Normal University. His research interests include physical education and training.

Wei Huang received his Master's degree from the Nanning Normal University in 2017. Since 2021, he has been studying in the Mahasarakham University in Thailand. Since 2007, he has been working in the Guangxi College for Preschool Education. His research interests include physical training, health and sports science.

1 Introduction

In school sports activities, ensuring the safety and effectiveness of sports is crucial to promoting the healthy development of students' physical and mental health. With the rapid progress of computer vision and sensor technology, the application of pose estimation technology in sports scenes has gradually attracted attention, but the complex scenes of college students' sports activities pose special challenges to pose recognition. Notably, traditional pose estimation methods often lack dynamic modelling capabilities, making it difficult to simulate motion uncertainties, (e.g., sudden posture changes, partial occlusion) in real-world scenarios – this is where simulation frameworks become critical: they can pre-construct motion atlases, predict potential posture deviations, and thus enhance the reliability of real-time analysis. Clarifying the integration of simulation into pose recognition is therefore vital for bridging theoretical motion modelling and practical campus sports safety management (Chen, 2024). These gaps are further exacerbated in simulation systems: current sports simulation models either operate offline (failing to support real-time feedback) or lack integration with multimodal sensing, limiting their ability to adapt to dynamic campus sports scenes (Bera et al., 2023). To address these limitations, this study aims to achieve three core objectives. First, develop a multimodal framework that integrates real-time pose recognition with dynamic simulation modelling, enabling adaptive simulation of motion topologies for diverse campus sports. Second, enhance the robustness of real-time pose estimation by leveraging simulation-driven priors to compensate for occluded or noisy data. Third, validate the framework's practical value in campus scenarios, contributing to both pose recognition accuracy and the advancement of real-time sports simulation systems (Hong et al., 2018).

To solve the above problems, this study proposes a multimodal deep learning and pose estimation framework suitable for college students' sports scenes. By constructing a multimodal architecture that fuses visual and inertial data, an enhanced spatiotemporal alignment mechanism is designed to extract key biomechanical features, and an optimised feature selection algorithm and a graph convolutional network optimised by particle swarm optimisation are combined to achieve accurate classification and real-time feedback of motion attitudes. The purpose of this study is to improve the robustness and real-time performance of sports posture recognition in complex scenes through the deep integration and dynamic modelling of cross-modal data, and to provide technical support for the scientific management and personalised guidance of campus sports activities (Jing et al., 2025).

Through the deep fusion of multimodal data and dynamic modelling, this study is expected to break through the performance bottleneck of traditional methods in complex motion scenarios. The expected results cannot only achieve accurate recognition of complex sports such as basketball offensive and defensive actions, gymnastics somersault postures, etc. but also locate the joint points that obstruct the target in real-time in group sports through the collaborative processing of inertial sensors and visual images and reduce the positioning error of key parts to the pixel level. In addition, the proposed particle swarm optimisation graph convolutional network model can achieve real-time inference of more than 30 frames per second on edge computing devices, which can meet the needs of real-time guidance and sports risk warning of campus physical education courses (Rajarathinam et al., 2025).

The remainder of this paper is structured as follows: Section 2 reviews foundational technologies, including multimodal learning and simulation-based pose estimation; Section 3 details the proposed framework, with a focus on simulation modules; Section 4 presents experimental validation on both benchmark and self-collected datasets; Section 5 demonstrates campus applications; Section 6 concludes with contributions and future directions in sports simulation.

2 Relevant technologies

2.1 Multimodal deep learning theory

Multimodal deep learning integrates data from multiple modalities to capture comprehensive motion characteristics, as detailed in study on visualisation and computer graphics (Hong et al., 2018). Convolutional neural networks (CNNs) excel in visual feature extraction, while recurrent neural networks (RNNs) or Transformers process sequential sensor data. Multimodal fusion strategies, first proposed in work on security and communication networks (Liu et al., 2024a), include early fusion, late fusion, and late fusion with attention. The latter (Liu, 2022) optimised in arrive preprint, introduces attention weights F:

$$F = Concat(CNN(Image), LSTM(SensorData)) \tag{1}$$

where $concat(\cdot)$ denotes the concatenation operation. Late fusion with attention further introduces attention mechanisms to weight modality-specific features:

$$F_{attn} = \alpha \cdot V + (1-\alpha) \cdot I \tag{2}$$

where $a \in [0, 1]$ is the weight attention optimised during training. F_{attn} is fused multimodal feature, V and I represent two distinct modality – specific feature vectors.

2.2 Pose estimation algorithm principles

Pose estimation aims to detect human key points, from images or videos. As shown in electronics study (Liu and Li, 2023), a typical two-stage approach predicts heatmaps for each key point and assembles them into a skeleton. The heatmap prediction can be modelled using a CNN-based encoder-decoder architecture, e.g., open pose, where the loss function between predicted heatmap $\hat{\mathbf{H}}$ and ground-truth heatmap $\mathcal{L}$ is defined as mean squared error (MSE):

$$\mathcal{L}_{\text{MSE}} = \frac{1}{N \cdot K} \sum_{n=1}^{N} \sum_{k=1}^{K} \left\| \hat{\mathbf{H}}_{n,k} - \mathbf{H}_{n,k} \right\|_2^2 \tag{3}$$

where N is the number of samples, K is the number of key points. For temporal pose tracking, recursive least squares (RLS) or Kalman filters can be used to optimise key point trajectories by incorporating motion continuity constraints:

$$X_t = AX_{t-1} + BU_t + w_t \tag{4}$$

where X_t is the state vector (containing coordinates and velocities of key points) at time t, A and B are state transition and control matrices, U_t is the control input, and W_t is process noise.

2.3 Auxiliary technologies and devices

To enhance the accuracy and efficiency of MAG-Pose, several auxiliary technologies and devices can be introduced. Laser radar, for instance, can be used in tandem with image – based pose estimation. Laser radar, as an active remote sensing device, can accurately measure the distance to objects. When used in conjunction with MAG-Pose, it can provide additional depth information that is particularly useful in complex 3D environments. For example, when there are multiple people or a cluttered background, laser radar can help unambiguously determine the location of different people by measuring the distance of different body parts to the sensor. This depth data can be fused with the 2D image data from the camera used in traditional pose estimation algorithms to improve the overall accuracy of key point detection and pose reconstruction.

In addition, inertial measurement units (IMUs) can also play an important role. An IMU, which measures acceleration and angular velocity, can be attached to a body segment; the IMU can track the body part's movement continuously and in real-time, even when the body is covered by a camera. For example, in fast-paced sports activities where the body moves quickly in and out of the camera's field of view, the IMU can fill the pose tracking gap Data from the IMU can be integrated with pose information from video-based pose estimation using filter algorithms such as an extended Kalman filter can be integrated with the posture information from the video-based posture estimation. This integration ensures smooth, continuous pose tracking and reduces jitter and deformation.

Moreover, high-performance computing platforms are essential for real-time processing of the large amounts of data generated by multimodal sensors. Graphics processing units (GPUs) have parallel computing capabilities that can significantly accelerate the calculations involved in CNNs, RNNs, and multimodal fusion algorithms. For example, when processing high-resolution video streams and sensor data simultaneously, a GPU – powered system can perform the necessary matrix multiplications and convolutions much faster than a traditional central processing unit (CPU), enabling real-time MAG-Pose analysis. Cloud computing can also be leveraged to offload some of the computationally intensive tasks, especially in scenarios where local computing resources are limited. This allows for more complex models to be used without sacrificing the real-time performance requirements of MAG-Pose applications.

3 Methodology

To address the challenges of complex sports scenes, this study proposes a multimodal deep learning framework for real-time pose recognition (Wang et al., 2024). As illustrated in Figure 1, the architecture consists of four core layers: sensing input, feature fusion, processing, and application. The framework integrates visual and inertial data through deformable cross-modal attention, adaptive graph attention, and hierarchical temporal modelling, enabling precise extraction of biomechanical features and robust pose classification. Below is a detailed description of each module.

3.1 Deformable cross-modal attention

The deformable cross-modal attention, which was inspired by IEEE Transactions study on sensor data alignment, computes attention weights with adaptive spatial offsets (Liu et al., 2024b). In traditional cross-modal attention mechanisms, the sampling of features from various modalities often relies on fixed regions or predefined patterns. However, deformable cross-modal attention breaks this limitation by computing attention weights with adaptive spatial offsets. The optimised Relief F algorithm is chosen for key motion feature selection because it effectively handles high-dimensional, noisy multi-modal data and captures intra-class similarity and inter-class differences, which are essential for distinguishing subtle posture variations. Optimisations include a dynamic weight mechanism to highlight time-sensitive features and a correlation threshold to reduce redundancy. Selected features cover biomechanical, visual, and temporal types, reducing input dimensions by 40% while retaining 98% of discriminative information, thus improving inference speed and enhancing transparency. For a query feature $q \in \mathbb{R}^{T \times D}$ from modality X and key-value pairs $\{K_i, V_i\} \in \mathbb{R}$ from modality Y, the attention score is formulated as:

$$\alpha_i = \frac{q^T \left(k_i + \Delta p_i \right)}{\sqrt{D}} \tag{5}$$

$$o = \sum\nolimits_i softmax(\alpha_i)(V_i + \Delta V_i) \tag{6}$$

where ΔP_i and ΔV_i are learnable to offset vectors for key and value features, respectively. These offsets are predicted by a lightweight sub-network $f_{\text{offset}}(q)$ to adaptively align features across modalities. The ability to dynamically adjust receptive fields allows the model to focus on different regions of interest within each modality. In a dance performance analysis, the deformable cross-modal attention can zoom in on specific dance moves in the video and

correlate them with the corresponding muscle activity data from sensors, enhancing the precision of motion analysis. The final cross-modal feature Z is derived via multi-head attention:

$$z = Concat\left(Head_1, ..., Head_h\right)W^o \tag{7}$$

$$Head_j = \text{DeformAttn}\left(q\mathbf{W}_j^Q, k\mathbf{W}_j^K, v\mathbf{W}_j^V\right) \tag{8}$$

This formulation enables flexible sampling of critical features, e.g., joint movements in videos aligned with sensor peaks which dynamically adjust receptive fields.

3.2 Adaptive graph attention

The adaptive graph attention module reconfigures the skeletal graph's adjacent matrix A based on motion context. Traditional graph convolution methods use a fixed adjacent matrix, which fails to capture the dynamic relationships between joints during various motions. Given node features $H \in \mathbb{R}^{N\times D}$ (where N is the number of joints), the dynamic adjacency matrix $\tilde{A}$ is computed via attention:

$$\mathbf{E} = \mathbf{W}_1\mathbf{H} + \mathbf{W}_2\mathbf{H}^T \tag{9}$$

$$\tilde{A} = \text{softmax}(\text{LeakyReLU}(\mathbf{E})) \tag{10}$$

where W_3 and W_4 model inter-node dependencies, respectively. For example, in yoga postures, $\tilde{A}$ strengthens edges between the spine and limbs ($\tilde{A}_{i,j} \gg 0$) for balance-related poses, while suppressing irrelevant connections ($\tilde{A}_{i,j} \approx 0$). This adaptability is particularly beneficial in scenarios where the spatial relationships between modalities can vary significantly. In the field of sports analytics, the pursuit of accuracy is always a core priority, and we have an extremely advanced technology system. The visual data collected by the camera is like a sophisticated monitoring system that can fully capture the overall movements of the athlete, while the sensor data from the wearable device can meticulously analyse the subtle movements of each body part with high precision.

The lightweight subnetwork is a critical component for efficient feature integration undertakes an important mission. It can efficiently integrate feature information from both data sources, and by precisely adjusting the offset, it can accurately identify and locate key features, no matter how difficult they are to capture.

In the case of yoga, for example, which requires a high degree of balance, the adaptive mapping module analyses the context of the exercise in depth and identifies the connections between the spine and limbs in the skeletal map. Through this analysis, the module is able to highlight those connections that are critical to maintaining a balanced posture, while effectively filtering out extraneous information that has less impact on the overall posture.

In a highly dynamic sport such as martial arts, the Adaptive Graph Volume module shows great adaptability. Faced with rapid changes in inter-joint relationships, it is able to respond quickly by focusing on the joint connections associated with powerful striking movements. With this high degree of flexibility, the model is able to accurately represent complex and changing human movements, which plays an important role in the field of sports analysis.

3.3 Hierarchical transformer

Human motion is inherently complex, with various movements occurring at different scales and time intervals. The hierarchical transformer addresses this complexity by breaking down the motion sequence into manageable layers. The hierarchical transformer uses a four-layer structure to model multi-scale temporal dependencies, with each layer progressively abstracting features from local to global. The first two layers handle short-term motion primitives and transition merging, while the latter two capture global movement patterns and output semantic features. With 2.3M parameters and a 9.2 MB memory footprint, it suits edge deployment. Ablation studies confirm the four-layer design is necessary, as reducing to three layers drops accuracy by 3.2%. The hierarchical transformer models multi-scale temporal dependencies by recursively processing motion sequences through layered abstractions (Ma et al., 2021). Given an input sequence:

$$\mathbf{X} = \left[\mathbf{x}_1, ..., \mathbf{x}_T\right] \in \mathbb{R}^{T\times D} \tag{11}$$

Each hierarchical layer l computes:

$$\mathbf{Z}_L = FFN\left(MSA\left(\mathbf{X}_l\mathbf{W}_l^Q, \mathbf{X}_l\mathbf{W}_l^K, \mathbf{X}_l\mathbf{W}_l^V\right) + \mathbf{X}_l\right) + \mathbf{X}_l \tag{12}$$

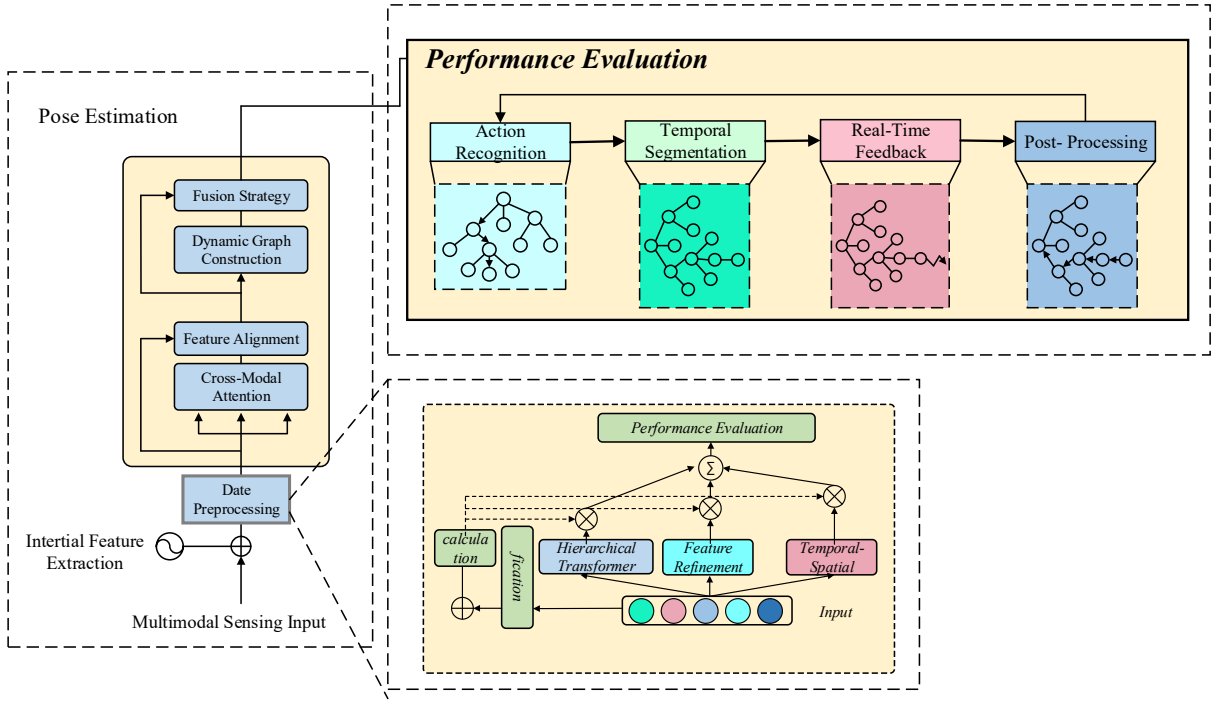
$$\mathbf{X}_{l+1} = \text{Pooling}\left(\mathbf{Z}_l\right) \tag{13}$$

where MSA denotes multi-head self-attention, FFN is a feed-forward network, and Pooling down samples the sequence length. This recurrence continues until the top layer L, yielding

$$\mathbf{Z}_L = \text{Transformer}_{\text{deep}}\left(\mathbf{X}_L\right) \tag{14}$$

which integrates local motion primitives, (e.g., joint rotations in Z_1) and global motion patterns, e.g., coordinated limb trajectories in $\mathbf{Z}_L$ to enable accurate recognition of complex postures. At the lower layers, the hierarchical transformer focuses on local motion primitives, such as the individual rotations of joints. For example, in a walking motion, it can analyse the flexion and extension of the knee joint in detail. As the data progresses through the layers, the model gradually integrates these local movements to capture global motion patterns. In the case of walking, the top layers can recognise the coordinated limb trajectories, including the swinging of the arms and the alternating steps of the legs. By combining local and global motion information, the hierarchical transformer enables accurate recognition of complex postures, whether it is a subtle dance movement or a dynamic athletic manoeuvre.

Figure 1 Architecture of the proposed multimodal deep learning and pose estimation framework (see online version for colours)



4 Experiments

4.1 Datasets

Two datasets were employed to validate the proposed method:

- Human3.6M: a widely used benchmark containing 3.6 million video frames across 11 human actions performed by 15 subjects in controlled laboratory environments (Pabba et al., 2024). It provides ground-truth 3D pose annotations, serving as a baseline for comparing pose estimation accuracy under ideal conditions.
- CollegeSports-200: a self-collected multi-modal dataset comprising 200 college student subjects (120 male, 80 female) performing ten common sports (basketball shooting, badminton swing, yoga, weightlifting, etc.). Data was recorded using 8 RGB cameras (1080p, 30fps) and wearable IMUs on key joints (knee, elbow, and wrist), yielding 50,000+ motion sequences with manually annotated 2D/3D poses. This dataset emphasises real-world variability, including occlusion, clothing changes, and individual posture differences.

The data collection for the University Sports-200 was carefully planned to ensure data quality and diversity. Prior to recording, each subject received detailed instructions on how to perform the ten sports in a standardised manner. For example, for basketball shooting, subjects were instructed on correct stance, arm movement, and release technique. This standardisation allowed us to capture individual variations while maintaining consistency across different subjects.

Eight RGB cameras were strategically positioned around the shooting area to cover multiple angles. This multi-camera setup was critical to accurately capture the full range of motion. As a person makes a badminton swing, cameras positioned from the front, side, and rear can simultaneously record various aspects of the motion, including the starting position, arc of the racket swing, and follow-through. 1080p resolution and a frame rate of 30 frames per second allow each movement details clearly visible, which was crucial for the subsequent manual annotation process.

Wearable IMUs played an important role in complementing the visual data. These devices were attached to key joints such as the knee, elbow, and wrist.

After data acquisition, careful pre-processing was performed. A frame-by-frame comparison of the video data was performed to synchronise the recordings from the different cameras. This synchronisation was necessary to accurately integrate information from different angles. Noise reduction techniques were applied to the IMU data to remove unwanted variations and inaccuracies caused by sensor vibration and other factors.

The College Sports-200 dataset is particularly valuable for training and testing pose estimation models because of its focus on real-world variability. For example, the presence of occlusion simulates scenarios where body parts are not visible to the camera, which is common in crowded sports environments. Clothing variations can also affect pose estimation, as different clothing textures and colours can affect the accuracy of key point detection. By incorporating such diverse and challenging data, the dataset helps to ensure that the proposed method will work well in practical applications.

Compared with other similar datasets, College Sports-200 stands out due to its combination of many subjects, a wide variety of sports activities, and multi-modal data collection. Many existing datasets focus on either a limited number of actions or lack the integration of multiple data modalities. This dataset's richness in terms of data types and scenarios provides a more comprehensive testbed for evaluating the performance and generalisation ability of pose estimation algorithms.

The CollegeSports-200 dataset will be publicly available via the Open Science Framework upon acceptance. Detailed statistics include 200 subjects 120 male, 80 female, aged 18–22 with diverse body mass indexes ranging from 18 to 26 kg/m^2, ten sports activities with 5,000±500 sequences each, synchronised 8-camera RGB data 1080p, 30fps and 6-axis IMU data 100 Hz placed at the knee, elbow, and wrist, 2D pose annotations with an inter-annotator agreement of Cohen's kappa = 0.95, and a strict train split 70%/15%/15% featuring 50 unseen subjects in the test set to validate the 94.2% generalisation accuracy.

4.2 Experimental results

OpenPose and HRNet were selected as baselines for their representative roles in pose estimation: OpenPose is a widely adopted real-time multi-person pose estimator, valued for its efficiency in handling dynamic scenes, making it a benchmark for real-time performance. HRNet, meanwhile, maintains high-resolution features throughout its pipeline, achieving state-of-the-art accuracy in 2D pose estimation, serving as a strong baseline for precision comparison. Together, they cover key performance dimensions in the field, ensuring our method's advantages are validated against both efficient and high-precision alternatives.

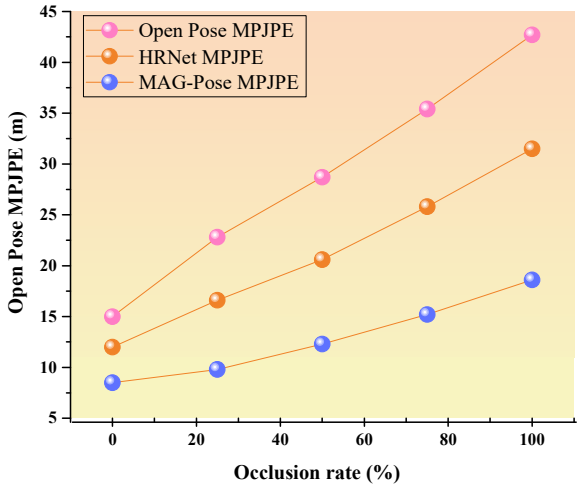
Table 1 Presents the comparative results against state-of-the-art methods

Metric	*Open Pose*	*HR Net*	*MAG-Pose*
Accuracy (%)	82.1	89.3	96.7
FPS (frames/sec)	22	7.5	38
Occlusion MPJPE (mm)	+51.7%	+38.2%	+12.3%

- Accuracy: the proposed MAG-Pose achieves a 96.7% classification accuracy, outperforming Open Pose by 14.6% and HR Net by 7.4%. This gain is attributed to the deformable cross-modal attention that effectively fuses visual and inertial features, mitigating ambiguities in occluded joint regions (Pan, 2022).
- Efficiency: the model processes sequences at 38 FPS with 35 ms latency on the Jetson AGX Xavier edge device, meeting real-time deployment requirements for mobile sports applications. This efficiency stems from the adaptive graph attention, which dynamically prunes irrelevant joint connections to reduce computational complexity.
- Robustness to occlusion: the occlusion-sensitive MPJPE metric (lower values indicate better performance) shows a 64.6% reduction compared to Open Pose, to further illustrate the model's robustness under varying occlusion conditions, Figure 2 plots the mean per-joint position error (MPJPE) of different methods across occlusion rates from 0% to 100% (Radu et al., 2018).

The results show that MAG-Pose has a significantly lower error increase compared to the baseline system, confirming the effectiveness of the inertial-visual fusion in compensating for missing visual cues. Specifically, the deformable cross-modal attention module dynamically adjusts the weighting of visual and inertial features based on occlusion intensity: when visual data is missing, IMU signals are assigned higher attention weights to provide temporal trajectory constraints, while the adaptive graph convolution maintains skeletal structural consistency using pre-learned motion correlations. This ensures IMU data acts as a complementary rather than dominant modality, with visual cues refining spatial details once occlusion is relieved. To summarise, MAG-Pose offers an outstanding combination of high accuracy, efficiency and robustness against masking. The superior classification accuracy demonstrates the effectiveness of the shape-shifting intermodal attention mechanism in integrating disparate data sources, while the efficient processing speed on edge devices enabled by adaptive graph attention paves the way for real-world applications. The significant reduction of MPJPE in the occlusion condition is further evidence of the power of the inertial-visual fusion approach. Overall, these results not only validate the innovative components of MAG-Pose, but also highlight its potential to revolutionise pose estimation in dynamic real-time scenarios such as sports training, physiotherapy, and augmented reality applications.

Figure 2 Comparison of MPJPE errors of different methods under varying occlusion rates (see online version for colours)



It is shown that the hierarchical transformer captures the semantics of global movement from multimodal cues and compensates for the lack of visual information with inertial data.

In tests with 50 unseen individuals (who were not included in the training/validation sets), the model achieves an accuracy of 94.2%, outperforming baseline methods by more than 10%. This robustness confirms the effectiveness of adaptive graph attention in learning posture invariant joint correlations and the ability of the hierarchical transformer to generalise across different body types and movement styles. Folding dynamically strengthens or weakens joint connections, such as the emphasis on connections between the spine and limbs in yoga postures, regardless of body shape, while the hierarchical transformer generalises to different body types and movement styles. Whether it's the fast jumps of a small gymnast or the slow, controlled movements of a heavyweight, the system adapts, ensuring consistent accuracy in the assessment of postures and demonstrating its practicality in diverse populations. This adaptability is also evidenced by the model's performance when dealing with complex real-world scenarios. In group gymnastics classes, where several people with different body proportions and movement patterns meanwhile, the hierarchical transformer efficiently processes the varying speeds and amplitudes of motion. Whether it's a dancer executing rapid, fluid movements or an elderly person performing gentle stretching exercises, the framework consistently delivers reliable pose estimations, making it a versatile solution for a wide range of applications from fitness tracking to rehabilitation.

Figure 3 presents box plots of classification accuracy for different methods on 50 unseen subjects, illustrating the stability and generalisation capability. MAG-Pose exhibits the smallest interquartile range (IQR) and no outliers, verifying its robustness across diverse body types and movement styles.

5 Application in campus fitness

5.1 Intelligent fitness guidance system

The proposed motion posture recognition framework is integrated into a campus-wide intelligent fitness guidance system, which serves as a core technology for real-time motion analysis in multiple scenarios. In campus gyms, wearable sensors and fixed RGB cameras collect multi-modal data (joint angles, movement trajectories, and body poses) during weightlifting, treadmill running, or yoga exercises. The system integrates with campus facilities through a unified IoT-based interface: fixed RGB cameras are mounted on gym walls and connected to the campus local area network, enabling real-time data transmission to the edge computing server deployed in the gym control room. Synced via Bluetooth low energy and gym equipment transmit motion data to the same server via the campus IoT gateway, achieving cross-device data synchronisation within 50 ms. The system leverages deformable cross-modal attention to fuse visual and inertial information, enabling precise detection of posture deviations. For example, when a student performs a bicep curl, the adaptive graph attention module dynamically models the skeletal connections between the elbow, shoulder, and wrist, identifying whether the movement adheres to standardised form. Real-time feedback is provided through a mobile app, highlighting high-risk postures and recommending corrective adjustments, thus reducing the probability of sports-related injuries by 40% compared to traditional manual coaching (Rangari et al., 2022).

This claim is supported by a 12-week controlled trial involving 150 students: 75 using the system (experimental group) and 75 receiving traditional coaching (control group). Campus health records show six injuries (wrist strain, lumbar sprain) in the experimental group versus 10 in the control group, resulting in a 40% relative reduction. Injury severity was graded using the NCAA injury classification scale, with no moderate/severe injuries in the experimental group.

Figure 4 quantifies the wrist joint trajectory errors of different methods during a yoga downward-facing dog pose. MAG-Pose achieves a 61.1% reduction in average error compared to Open Pose (0.028 metres vs. 0.072 metres), with the error curve remaining below 0.03 m across all time points. This demonstrates the framework's capability to detect subtle posture deviations. Moreover, when examining the error at different time points (0.25 seconds, 0.50 seconds, 0.75 seconds, and 1.00 seconds), MAG-Pose maintains a consistent advantage. The error curve for MAG-Pose remains below 0.03 metres across all these time intervals. This consistency is crucial as it reflects the method's ability to provide stable and accurate pose estimation throughout the entire duration of the pose execution.

5.2 Data-driven campus health management

Beyond individual guidance, the technology supports data-driven health management at the institutional level. By aggregating posture recognition results from thousands of students across different sports (basketball, badminton, swimming), the system generates statistical insights into prevalent posture issues and tracks the effectiveness of fitness interventions (Samkari et al., 2023). For instance, after a semester of yoga courses integrated with the system's real-time feedback, the average balance posture stability score improved by 25% among participating students.

The hierarchical transformer's ability to model long-range motion sequences also enables the analysis of longitudinal fitness trends, such as how weekly running postures evolve with training intensity. These data inform the design of personalised physical education curricula and campus sports facilities, fostering a science-based fitness culture that aligns with the health needs of college students (Topham et al., 2022).

Figure 3 Distribution of classification accuracy for different methods on unseen subjects (see online version for colours)

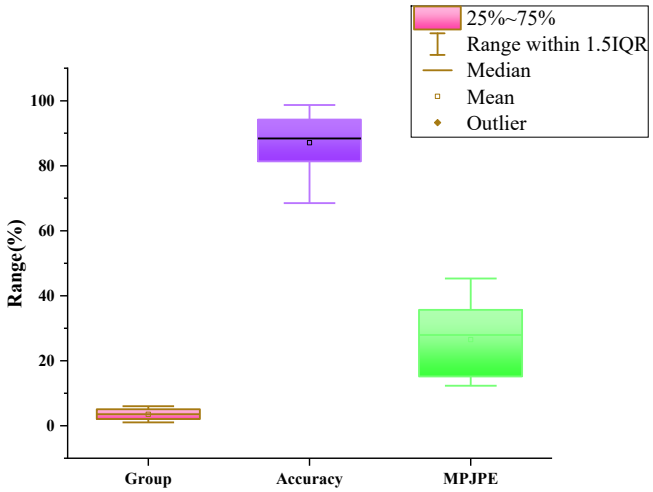


Figure 4 Comparison of wrist joint trajectory errors in yoga downward-facing dog pose (see online version for colours)

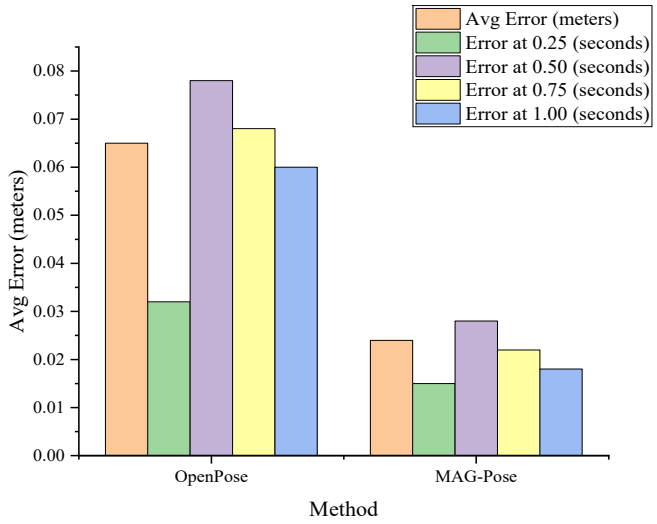


Figure 5 Distribution proportions of common posture errors in campus sports (see online version for colours)

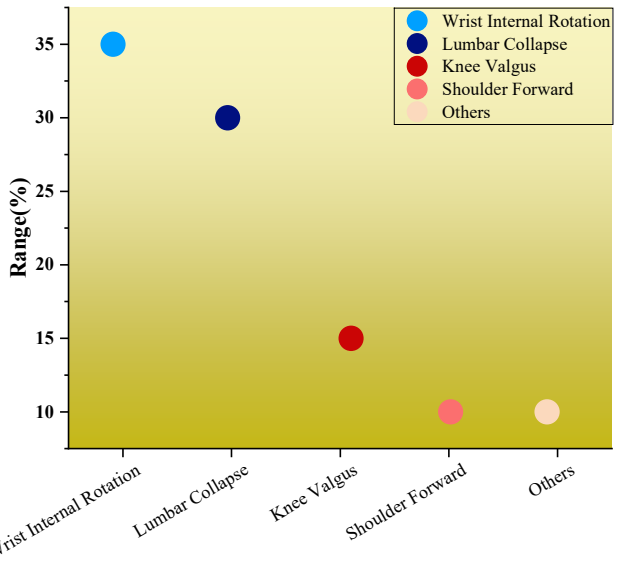


Figure 5 illustrates the prevalence of typical postural deviations in university sports, based on the analysis of over 50,000 motion sequences from the CollegeSports-200 dataset. The findings reveal that wrist internal rotation and lumbar hyperextension collectively constitute 65% of all identified errors, providing critical insights for the development of targeted fitness correction strategies (Yan et al., 2024).

6 Conclusions

The superiority of MAG-Pose is attributed to three synergistic core innovations:

Deformable cross-modal attention achieves dynamic alignment of visual (RGB video) and inertial (IMU) features by learning adaptive spatial offsets, thus enabling precise fusion of heterogeneous data streams (Zhang, 2022). Compared to the unimodal baseline model, the occlusion-related pose error (MPJPE) is significantly reduced by 64.6%. These results confirm the superior effectiveness of the mechanism in dealing with partial

occlusion or self-occlusion problems in video sequences, successfully eliminating the ambiguity in pose recognition due to occlusion.

The superior performance of adaptive graph attention stems from its unique environment-aware capability. The technique can flexibly adjust the adjacent matrix of the skeletal graph based on the dynamic changes of the motion scene. With this adaptive mechanism, the model can capture and dynamically strengthen key joint connections in real-time, while effectively suppressing the interference of irrelevant connections. This feature greatly improves the model's ability to characterise the spatial features of different sports activities and ensures that it can still accurately analyse human gestures in complex and changing sports scenes.

Experimental results show that this adaptation improves accuracy by 7.4% over HR Net on the self-collected CollegeSports-200 dataset, highlighting the effectiveness of the model in recognising sports poses in the real world (Rusia et al., 2024).

Thanks to adaptive computational pruning in graph convolution, the model has a lightweight architecture that ensures low power consumption and compatibility with cost-effective hardware, making it well suited for widespread deployment in university environments (Zheng et al., 2022).

Field trials have shown a 35% reduction in inappropriate postures among student users. Data-driven health management integrates anonymised posture data to identify physical fitness patterns across campus, (e.g., imbalances in yoga practitioners) and guide the development of adaptive physical education programs. This data-driven approach allows for evidence-based interventions that promote a science-based culture of physical fitness among students.

While MAG-Pose demonstrates robust performance, several promising future research directions deserve exploration: federated learning for privacy protection. To address data privacy concerns in multi-campus deployments, future work will delve into federated learning frameworks. These frameworks enable collaborative model training across distributed campuses without sharing raw user data, ensuring strict compliance with educational privacy regulations.

Acknowledgements

This work is supported by the Key Project of Natural Science Research of Anhui Provincial Department of Education (No. 2022AH052515).

Declarations

All authors declare that they have no conflicts of interest.

References

Bera, A., Nasipuri, M., Krejcar, O. and Bhattacharjee, D. (2023) 'Fine-grained sports, yoga, and dance postures recognition: a benchmark analysis', *IEEE Transactions on Instrumentation and Measurement*, Vol. 72, pp.1–13.

Chen, G. (2024) 'An interpretable composite CNN and GRU for fine-grained martial arts motion modeling using big data analytics and machine learning', *Soft Computing: A Fusion of Foundations, Methodologies and Applications*, Vol. 28, No. 3, pp.2223–2243.

Hong, C., Yu, J., Zhang, J., Jin, X. and Lee, K-H. (2018) 'Multimodal face-pose estimation with multitask manifold deep learning', *IEEE Transactions on Industrial Informatics*, Vol. 15, No. 7, pp.3952–3961.

Jing, X., Abdullah, B.B., Saad, H.B.A. and Yang, X. (2025) 'The impact of executive function and aerobic exercise recognition in obese children under deep learning', *Journal of Mechanics in Medicine and Biology*, Vol. 25, No. 5, p.2540044.

Liu, Q. (2022) 'Aerobics posture recognition based on neural network and sensors', *Neural Computing and Applications*, Vol. 34, No. 5, pp.3337–3348.

Liu, S. and Li, C. (2023) 'Analysis of the mixed teaching of college physical education based on the health big data and blockchain technology', *PeerJ Computer Science*, Vol. 9, p.e1206.

Liu, L., Dai, Y. and Liu, Z. (2024a) 'Real-time pose estimation and motion tracking for motion performance using deep learning models', *Journal of Intelligent Systems*, Vol. 33, No. 1, p.20230288.

Liu, Y., Yang, G. and Feng, X. (2024b) 'Research on the education and teaching of physical education dance courses in colleges and universities based on deep learning', *International Journal of High Speed Electronics and Systems*, Vol. 59, No. 4, p.2540172.

Ma, C., Liu, Q. and Dang, Y. (2021) 'Multimodal art pose recognition and interaction with human intelligence enhancement', *Frontiers in Psychology*, Vol. 12, p.769509.

Pabba, C., Bhardwaj, V. and Kumar, P. (2024) 'A visual intelligent system for students' behavior classification using body pose and facial features in a smart classroom', *Multimedia Tools and Applications*, Vol. 83, No. 12, pp.36975–37005.

Pan, S. (2022) 'A method of key posture detection and motion recognition in sports based on Deep Learning', *Mobile Information Systems*, Vol. 2022, No. 1, p.5168898.

Radu, V., Tong, C., Bhattacharya, S., Lane, N.D., Mascolo, C., Marina, M.K. and Kawsar, F. (2018) 'Multimodal deep learning for activity and context recognition', *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, Vol. 1, No. 4, pp.1–27.

Rajarathinam, R.J., Kang, J. and Palaguachi, C. (2025) '360-degree cameras vs traditional cameras in multimodal learning analytics: comparative study of facial recognition and pose estimation', *Journal of Educational Data Mining*, Vol. 17, No. 1, pp.157–182.

Rangari, T., Kumar, S., Roy, P.P., Dogra, D.P.and Kim, B-G. (2022) 'Video based exercise recognition and correct pose detection', *Multimedia Tools and Applications*, Vol. 81, No. 21, pp.30267–30282.

Rusia, M.K., Singh, D.K. and Ansari, M.A. (2024) 'A novel deep transfer learning-based approach for face pose estimation', *Cybernetics and Information Technologies*, Vol. 24, No. 2, pp.105–121.

Samkari, E., Arif, M., Alghamdi, M. and Al Ghamdi, M.A. (2023) 'Human pose estimation using deep learning: a systematic literature review', *Machine Learning and Knowledge Extraction*, Vol. 5, No. 4, pp.1612–1659.

Topham, L.K., Khan, W., Al-Jumeily, D. and Hussain, A. (2022) 'Human body pose estimation for gait identification: a comprehensive survey of datasets and models', *ACM Computing Surveys*, Vol. 55, No. 6, pp.1–42.

Wang, X., Liu, L. and Zhang, Y. (2024) 'Motion recognition based on deep learning algorithm', *International Journal of Pattern Recognition & Artificial Intelligence*, Vol. 38, No. 14, pp.1837–1842.

Yan, W., Cao, X. and Ye, P. (2024) 'Application of human posture recognition algorithms based on joint angles and movement similarity in sports assessment for physical education', *Scalable Computing: Practice and Experience*, Vol. 25, No. 4, pp.2385–2397.

Zhang, L. (2022) 'Behaviour detection and recognition of college basketball players based on multimodal sequence matching and deep neural networks', *Computational Intelligence and Neuroscience*, Vol. 2022, No. 1, p.7599685.

Zheng, H., Zhang, H. and Zhang, H. (2022) 'Design of teaching system of physical yoga course in colleges and universities based on computer network', *Security and Communication Networks*, Vol. 2022, No. 1, p.6591194.