SIN-CHN
SCIENTIFIC PRESS PTE.LTD.

Molecular & Cellular Biomechanics 2025, 22(4), 1581.
https://doi.org/10.62617/mcb1581

Article

Transformer-based video generation technique for biomechanical motion analysis

Yuxuan Jia

The Central Academy of Drama, Beijing 100710, China; yuxuanjiacad@163.com

CITATION

Jia Y. Transformer-based video generation technique for biomechanical motion analysis. Molecular & Cellular Biomechanics. 2025; 22(4): 1581. https://doi.org/10.62617/mcb1581

ARTICLE INFO

Received: 17 February 2025
Accepted: 5 March 2025
Available online: 14 March 2025

Abstract: In this study, a Transformer-based video generation technique is proposed for accurately modelling biomechanical movement patterns, and its performance is systematically evaluated in walking, running, throwing and other movement tasks. The experimental results show that Transformer outperforms traditional methods (RNN, CNN, GAN) in terms of motion trajectory consistency, temporal synchronization, and video clarity, and is capable of generating high-quality motion videos that comply with biomechanical constraints. This study not only expands the application scope of Transformer in biomechanical analyses, but also provides high-precision solutions for tasks such as gait reconstruction, abnormality detection, rehabilitation training, and motion prediction.

Keywords: biomechanical motion analysis; transformer; video generation; motion prediction

1. Introduction

Biomechanical motion analysis has important application value in the fields of sports science, rehabilitation medicine, and robot control, etc. Traditional motion analysis methods mainly rely on motion capture systems, inertial sensors (IMUs), and computer vision for data acquisition. However, these methods have limitations such as high cost of data annotation, high computational complexity, and strong dependence on the environment, making accurate motion prediction and reconstruction tasks challenging. In recent years, the rapid development of deep learning and computer vision technologies has pushed forward data-driven motion analysis methods, among which Transformer provides a new solution for biomechanical motion analysis due to its powerful spatio-temporal feature modelling capability, which is superior in long-time dependent modelling, trajectory prediction and video generation.

2. Relevant technological foundations

2.1. Computer vision and video generation

The rapid development of computer vision in the field of video generation has opened up new possibilities for modelling complex motion sequences and biomechanical motion analysis. Traditional video generation methods rely on Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs), where CNNs excel at spatial feature extraction, while RNNs and their variants (e.g., LSTMs) are used to model time-series information [1]. These methods have limitations in modelling long time-series dependencies, making it difficult to accurately capture the detailed variations in biomechanical motion. In recent years, generative methods based on autoregressive models (e.g., Video PixelCNN), generative adversarial networks (GANs), and variational autoencoders (VAEs) have made significant progress in improving video quality, but still face the problems of coupled spatio-temporal information and insufficient long-distance-dependent capturing capability.

With the introduction of the self-attention mechanism, Transformer-based video generation technology has become an important breakthrough in this field. Compared with traditional methods, the transformer's global feature modelling capability enables it to effectively capture the long-term dependencies of motion trajectories, thus improving the spatio-temporal consistency and structural integrity of video generation. In biomechanical motion analysis, video generation techniques need to ensure not only the realism of the visual appearance, but also compliance with the laws of kinematics and dynamics [2]. Transformer-based video generation can establish global correlations between multiple time steps, and improve the modelling ability of complex motion patterns through motion state encoding and constraint learning, laying the foundation for accurate and efficient biomechanical motion analysis.

2.2. Application of transformer model in video generation

While Transformer models offer superior long-range dependency modelling capabilities, they come with increased computational complexity. Processing high-dimensional video data with self-attention mechanisms requires substantial memory and processing power, making real-time applications challenging. To address this, several optimization strategies have been explored:

Sparse Attention Mechanisms: Reducing the quadratic complexity of standard self-attention by focusing on local spatio-temporal dependencies.

Factorized Space-Time Attention: Splitting attention computation into spatial and temporal components to reduce overhead [3].

Efficient Transformer Variants: Implementations such as Swin Transformer and Linformer introduce hierarchical structures and linearized attention to enhance efficiency.

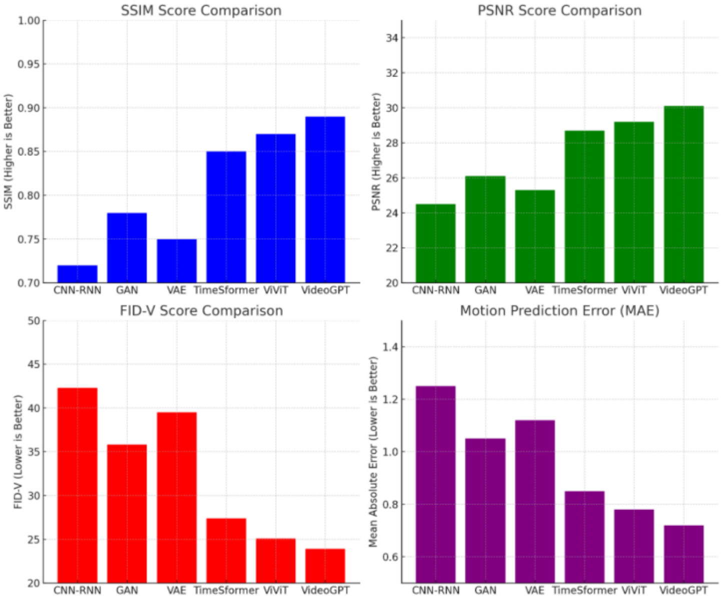

Figure 1. Comparison of video generation performance of transformer.

However, despite these optimizations, deploying Transformer-based video generation on edge devices or real-time applications remains a challenge, requiring future work on lightweight model architectures and hardware acceleration, such as **Figure 1**.

While Transformer-based methods excel in long-term motion prediction and biomechanical consistency, they may struggle with short-duration rapid movements. Sudden, high-velocity changes, such as throwing or abrupt shifts in direction, introduce motion discontinuities that are difficult to model accurately with self-attention mechanisms. Experimental results indicate that: Walking and Running Tasks: Transformer achieves high trajectory consistency due to its long-range feature modelling [4]. Throwing and Sudden Motion Tasks: GAN-based approaches tend to produce more natural and flexible motion sequences, although they may sacrifice biomechanical accuracy. Future work should explore hybrid models combining Transformer’s temporal modelling strength with GAN’s adaptability for short-term movements.

2.3. Biomechanical motion analysis

Biomechanical motion analysis aims to study the dynamics and kinematic properties of human motion, and to resolve the mechanical behavior of the motion system in different environments through mathematical modelling, experimental measurements and computer simulations [5]. With the support of computer vision and deep learning, biomechanical research has expanded from traditional motion capture techniques (e.g., optical marker points, inertial sensors) to video-based depth estimation methods, enabling contactless and non-invasive human motion analysis. In particular, in Transformer-based video generation tasks, biomechanical motion analysis is not only used for data annotation and motion pattern extraction, but also for improving the motion rationality and temporal consistency of the generated video through kinetic constraints. In terms of mathematical modelling, biomechanical motion analysis usually adopts the Inverse Dynamics method to calculate the joint forces with the following formula:

$$\tau = J^T(F_{ext} - mg - m\ddot{x}) \quad (1)$$

where τ denotes the joint moment, J is the Jacobi matrix, F_{ext} is the external force, m is the body mass, g is the gravitational acceleration, and $\ddot{x}$ is the acceleration vector. This formulation is used to estimate the forces on an individual in different motion states and combined with Transformer for time-series modelling to improve the physical consistency of motion generation. Human motion can be modelled as a Multi-body System (MBS) and its trajectory can be described by the Lagrangian Dynamics equations:

$$\frac{d}{dt}\left(\frac{\partial L}{\partial \dot{q}_i}\right) - \frac{\partial L}{\partial q_i} = Q_i \quad (2)$$

where $L = T - V$ is the Lagrangian quantity, q_i is the generalized coordinate, T is the kinetic energy, V is the potential energy and Q_i is the generalized force. This equation can be used to calculate the energy conversion process under different motion modes and combined with Transformer to predict the motion state of future frames, thus generating video sequences that conform to biomechanical laws. Combined with Transformer's global temporal feature learning capability, biomechanical motion analysis can achieve more accurate motion prediction, anomaly detection and video generation, and promote the in-depth development of intelligent motion analysis and simulation [6].

3. Transformer-based video generation technology

3.1. Transformer architecture for video generation

The proposed method utilizes masked modelling and contrastive learning to improve motion sequence reconstruction [7]. However, large-scale datasets are essential for training robust Transformer models. The reliance on vast labeled data increases training costs, as biomechanical video datasets require precise annotation through motion capture systems or manually labeled skeletal sequences. The key to Transformer's processing of video is how to efficiently decouple spatio-temporal information and reduce computational complexity. reduce computational complexity. Common methods include:

(1) Factorized Space-Time Attention: In the Transformer structure, spatial and temporal attention are separated to reduce computational overhead:

$$A_{spatial} = Soft\,max\left(\frac{Q_s K_s^T}{\sqrt{d_k}}\right)V_s \tag{3}$$

$$A_{temporal} = Soft\,max\left(\frac{Q_t K_t^T}{\sqrt{d_k}}\right)V_t \tag{4}$$

where Q, K, V denote Query, Key, Value matrices respectively, and d_k is the dimensional scaling factor.

(2) Divided Space-Time Attention: Divides the video sequence into different time windows, performs the attention computation in the local area, and effectively reduces the cost of Transformer computation:

$$\mathrm{Z} = \mathrm{SelfAttention\ (PatchEmbed(X))} \tag{5}$$

where X is the input video frame, PatchEmbed is the projection module for feature extraction, and Z represents the spatio-temporal features extracted by the Transformer.

(3) Masked Video Modeling (Masked Video Modeling): Similar to BERT's Masked Language Model (MLM), some video frames are randomly masked, and the model is required to speculate the missing content based on the known frames, which improves the model's temporal modelling capability [8].

In specific implementations, ViViT (Video Vision Transformer) uses hierarchical spatio-temporal feature extraction, TimeSformer uses independent spatial-temporal attention, and VideoGPT uses an autoregressive video generation strategy. These methods perform well in biomechanical applications such as motion prediction, frame completion, and anomaly detection, such as **Figure 2**.

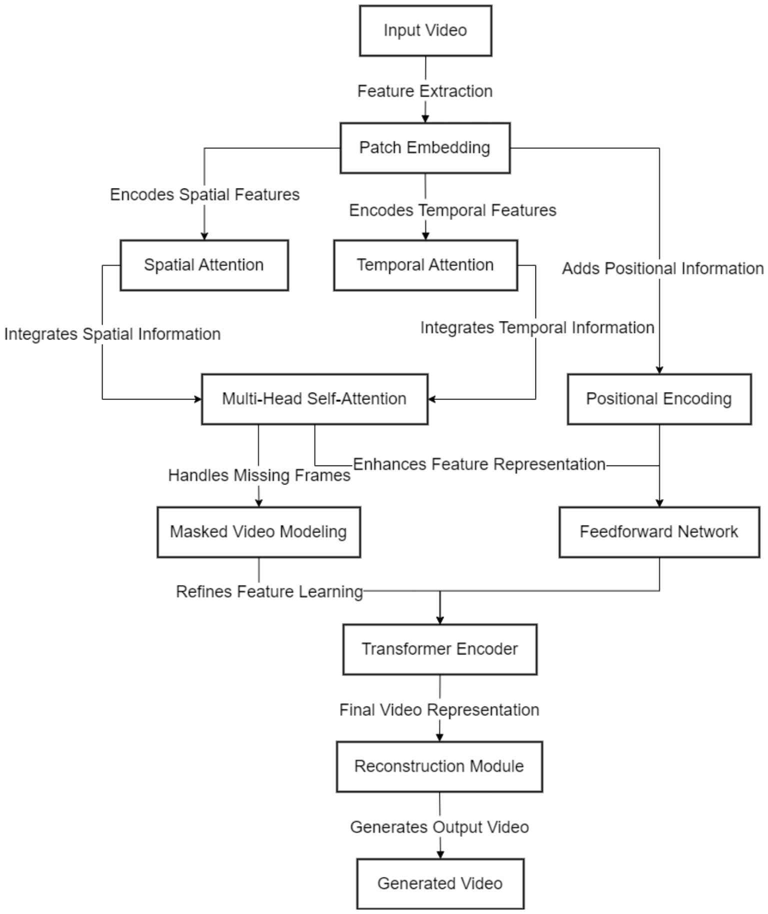

Figure 2. Overall architecture of the transformer in the video generation task.

The Transformer-based video generation architecture achieves efficient video sequence modelling by decoupling spatial and temporal features. The input video is subjected to Patch Embedding for feature extraction, and spatial and temporal features are modelled by Spatial Attention and Temporal Attention, respectively. Multi-Head Self-Attention combines the information from different attention heads to enhance the spatio-temporal feature fusion capability [9]. Masked Video Modeling introduces frame-missing training to improve the motion prediction ability of the model. After Transformer Encoder encoding and Reconstruction Module processing, the video sequences are synthesized to conform to the physical constraints, providing accurate prediction and reconstruction capabilities for biomechanical motion analysis.

3.2. Temporal feature modelling for motion analysis

Temporal feature modelling for motion analysis mainly involves dynamics modelling of human motion sequences, temporal dependency capturing and non-linear motion pattern learning. In the video generation task, the timing information determines the coherence of the movements and also affects the reasonableness of the motion trajectories [10]. Traditional sequence modelling methods such as Long Short-Term Memory Networks (LSTMs) and Temporal Convolutional Networks (TCNs) are able to extract temporal features to a certain extent, but they are difficult to model long-distance dependencies and do not make full use of the global information. The Transformer's Self-Attention mechanism enables the model to learn motion patterns throughout the entire video sequence by associating global features. video sequences to learn motion patterns, thus improving the quality of motion prediction and video generation. In terms of mathematical modelling, human motion can be represented as a discrete time sequence $X = \{x_1, x_2, \ldots, x_T\}$, where x_t represents the motion state at time *t*. Usually, we need to model the dynamic changes of the motion state in the time dimension, i.e., to solve the state transfer function:

$$x_{t+1} = f(x_t, u_t) + \varepsilon_t \tag{6}$$

where $f(x_t, u_t)$ denotes the state transfer function, u_t is a control input, and ε_t is a perturbation term. In the deep learning model, this state transfer function can be modelled by Transformer, where the self-attention mechanism is calculated as follows:

$$Attention(Q, K, V) = Soft\,max\left(\frac{QK^T}{\sqrt{d_k}}\right)V(7) \tag{7}$$

where Q, K, V denote the Query, Key and Value matrices respectively and d_k is the dimensional scaling factor. The mechanism is able to compute the weighting relationship between different time steps to capture complex motion dependency patterns.

(1) Motion Feature Encoding and Temporal Dependency Modelling

In the Transformer structure, temporal features are usually augmented by Positional Encoding (Positional Encoding) to compensate for the lack of time-awareness of the self-attention mechanism. Positional encoding is defined as follows:

$$PE(i, 2i) = sin\left(\frac{t}{10000^{2i/d}}\right), \quad PE(t, 2i + 1) = cos\left(\frac{t}{10000^{2i/d}}\right) \tag{8}$$

where *t* is the time step, *d* is the feature dimension, and i is the index. This encoding enables the Transformer to identify the relative positions between time steps so that temporal information is not lost when modelling temporal dependencies.

(2) Motion trajectory prediction and time series regression

Motion trajectory prediction is an important part of biomechanical motion analysis, with the goal of predicting future motion states based on information from past frames. Mathematically, trajectory prediction can be expressed as a sequence regression problem, i.e., predicting a future trajectory when the past trajectory $\{x_1, \ldots, x_T\}$ is known $\{x_{T+1}, \ldots, x_{T+H}\}$. This task can be modelled by the Transformer decoder, and the objective optimization function is usually a mean square error (MSE):

$$L_{MAE} = \frac{1}{H}\sum_{h=1}^{H} \|x_{T+h} - \hat{x}_{T+h}\|^2 \tag{9}$$

where x_{T+h} is the true trajectory point and $\hat{x}_{T+h}$ is the predicted trajectory point. This loss function measures the prediction trajectory and true trajectory error and can optimise the motion sequence generated by Transformer to be more physically correct.

(3) Modelling of Motion Constraints and Kinetic Consistency

In biomechanical motion analysis, motion must be consistent with kinetic constraints such as velocity, acceleration, joint forces and other physical rules. Physical constraint-based loss functions can be introduced in the Transformer structure, for example:

$$L_{physics} = \sum_{t=1}^{T} \|M\ddot{x}_t - F_t\|^2 \tag{10}$$

where M is the mass matrix, $\ddot{x}_t$is the acceleration, and F_t is the force vector. This constraint ensures that the generated video motion trajectory conforms to the laws of physics and improves the realism of the motion prediction.

(4) Multimodal motion fusion and long time series modelling

Human motion analysis usually involves multiple modal data, such as video frames, skeletal points, electromyographic data (EMG), ground reaction force (GRF), and so on. In order to integrate these data, Multimodal Transformer (MT) can be used, whose core calculation formula is as follows:

$$H = Concat\left(H_{video,}H_{skeleton,}H_{EMG,}H_{GRF}\right) \tag{11}$$

$$MultiModalAttention(Q,K,V) = Soft\,max\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{12}$$

where $H_{video}, H_{skeleton}, H_{EMG}, H_{GRF}$ denotes video, skeletal, EMG and ground reaction force features, respectively, and MultiModal Attention calculates cross-modal correlations between them to improve motion prediction accuracy.

Modelling temporal features for motion analysis is a core aspect of video generation, which is crucial to capture temporal dependencies, optimize trajectory prediction and ensure physical consistency. Transformer models long-range motion dependencies through a self-attention mechanism and combine positional coding, lossy trajectory prediction and kinetic constraints to make the generated video sequences more accurate.

3.3. Motion data and transformer input mapping

Motion data usually includes a variety of modal information such as raw video frames, skeletal keypoints, electromyographic data (EMG), accelerometer data, and ground reaction force (GRF). These data need to be processed through feature extraction, embedding representation, and temporal modelling to fit the input requirements of the Transformer structure [11]. Mathematically, the motion data are assumed to be a time series$X = \{x_1, x_2, \ldots, x_T\}$ where x_t represents the motion state vector at time *t*. The goal is to map them to the Transformer's high-dimensional input representation H. This process consists of the following main steps:

(1) Motion Data Feature Extraction and Coding

Motion data usually come from multiple sensors or computer vision systems, and need to be preprocessed to extract effective features. For example, a sequence of skeletal keypoints based on attitude estimation can be used to capture human motion trajectories, and its state vector is defined as follows:

$$x_t = \{p_t^1, p_t^2, \ldots, p_t^J\} \tag{13}$$

where $p_t^i = \{x_t^j, y_t^j, z_t^j\}$ denotes the spatial coordinates of the *j*-th skeletal joint point, and there are a total of *J* key points. In order to unify the data format, a linear transformation is often used to normalize the data:

$$\tilde{x}_t^j = \frac{x_t^j - \mu}{\sigma}, \quad \tilde{y}_t^j = \frac{y_t^j - \mu}{\sigma}, \quad \tilde{z}_t^j = \frac{z_t^j - \mu}{\sigma} \tag{14}$$

where μ, σ are the mean and standard deviation of the data, respectively, to ensure stable data distribution and improve the convergence of model training. For electromyography data (EMG) or acceleration data (IMU), wavelet transform or Fourier transform can be used to extract the frequency domain features, so that they contain richer time series information:

$$X_{freq} = F(X) \tag{15}$$

where F(·) denotes the Fourier transform, which is used to convert the signal to the frequency domain, enabling the Transformer to learn the time-frequency characteristics of the motion pattern.

(2) Motion Data Embedding and Transformer Input Mapping

After feature extraction, the raw motion data needs to be mapped into the input space of the Transformer. Since the Transformer uses a fixed dimension input format, we need to perform Linear Projection on the motion sequence to adjust the dimension:

$$H_t = WX_t + b(16) \tag{16}$$

where $W \in R^{d_h \times d_x}$ is the learnable weight matrix, b is the bias, H_t is the high-dimensional feature representation of time step t, d_x is the dimension of the input motion data, and d_h is the input dimension expected by the Transformer. In practice, a Patch Embedding method can be used to divide the long-time motion sequence into multiple time windows P_t and then project it:

$$H = W \cdot Concat(P_1, P_2, \dots, P_T) \tag{17}$$

This improves the model's ability to learn local motion patterns and reduces computational complexity.

(3) Positional Encoding and Timing Information Enhancement

Since the Transformer structure is not time-aware, Positional Encoding needs to be introduced to embed timing information. Common encoding methods include Sinusoidal Encoding and Learnable Embedding. The mathematical expression for Sinusoidal-Cosine Encoding is as follows:

$$PE(i, 2i) = sin\left(\frac{t}{10000^{2i/d}}\right), \quad PE(t, 2i+1) = cos\left(\frac{t}{10000^{2i/d}}\right) \tag{18}$$

where *t* is the time step, *d* is the feature dimension, and *i* is the channel index. This method enables the model to learn the relative relationship between time steps, ensuring the continuity of the motion trajectory. For more complex motion data, such as multimodal sensor data fusion, learnable positional embedding, i.e., learning a separate vector for each time step, can be used:

$$H_t^{input} = H_t + PE_t \tag{19}$$

where PE_t is a learnable parameter that can be optimized for different tasks to improve the adaptability of timing modelling.

(4) Motion Data Time Windowing and Transformer Input Optimization

In order to improve the computational efficiency, the long time series data is usually processed by Sliding Window, i.e., the original data is divided into multiple sub-sequences and the attention computation is performed in a local window. Let the window size be W, then the input data of each window is:

$$X_{win} = \{x_t, x_{t+1}, \dots, x_{t+W}\} \tag{20}$$

The Transformer performs Self-Attention computation only on the data within the window to reduce the computational complexity and improve the learning ability of local motion features. A Factorized Attention approach can be used to split the temporal and spatial attention processing:

$$A_{space-time} = A_{space} \cdot A_{time} \tag{21}$$

Among them, A_{space} calculates the feature relationships in the spatial dimension and A_{time} calculates the feature relationships in the temporal dimension, and this approach effectively reduces the computational cost and improves the scalability of the model.

The key to mapping motion data to Transformer inputs lies in multimodal feature extraction, feature embedding, position encoding and time windowing. Motion data usually includes video frames, skeletal keypoints, EMG signals, accelerometer data, etc., which need to be normalized and frequency domain transformed to ensure the resolvability of the data [12]. Subsequently, the motion data is mapped to a high-dimensional input representation of the Transformer by means of linear projection and Patch Embedding. In order to enhance the time-dependent modelling capability, position coding and time windowing mechanisms are introduced to optimize the long sequence modelling effect.

4. Applications in biomechanical motion analysis

4.1. Motion prediction and reconstruction

Motion prediction and reconstruction is of great importance in biomechanical research and is widely used in the fields of sports rehabilitation, ergonomics, robot control and computer animation [13]. The core goal of this task is to analyze and learn human motion patterns, predict future trajectories from historical motion data, and reconstruct accurately in the presence of missing data or noise interference. With the support of Transformer and deep learning technologies, the accuracy and stability of motion prediction and reconstruction have been significantly improved, especially in the long time-series dependent modelling and high-dimensional motion data processing.

The core of motion prediction lies in the speculation of future states based on existing motion trajectories, which is crucial in tasks such as gait analysis, motion planning, and posture correction. Traditional prediction methods, such as Kalman Filtering, Markov Models and Dynamic Time Warping (DTW), mainly rely on linear extrapolation or probabilistic modelling, which are limited when facing complex biomechanical movement patterns. In recent years, deep learning methods (e.g., Transformer, LSTM, TCN) have been able to effectively extract long-term dependencies in motion sequences with the help of the global attention mechanism

(Self-Attention) to improve the accuracy and stability of motion prediction. Motion prediction input data usually include joint angles, velocities, accelerations, muscle activation signals, etc., which can be captured by inertial measurement units (IMUs), motion capture systems (MoCap), and video pose estimation (Pose Estimation). Compared with traditional methods, the advantages of the Transformer structure in motion prediction are: (1) Global information modelling, which improves the understanding of complex motion patterns by learning long-time dependencies through the self-attention mechanism; (2) adaptive learning, which improves prediction accuracy by personalizing the modelling for different individuals; and (3) non-linear modelling capability, which can effectively deal with irregular motion patterns and avoid the limitations of linear methods [14].

The core challenges of motion reconstruction include: (1) Spatial consistency to ensure that the reconstructed motion data conforms to biomechanical constraints, such as joint angle ranges, velocity stability, and muscle mechanical properties; (2) temporal continuity to avoid abrupt or unreasonable state jumps in the reconstructed motion sequences; and (3) multimodal fusion, which integrates the information between different data sources (IMUs, videos, force sensors, etc.) , to improve the accuracy of motion reconstruction. Modern deep learning techniques, such as Generative Adversarial Networks (GANs), Autoregressive Models (ARs), Variational Autocoders (VAEs), etc., are able to efficiently fill in the missing motion trajectories through nonlinear modelling. Transformer combines sequential modelling and self-supervised learning approaches to demonstrate higher stability and generalization capabilities in motion reconstruction tasks.

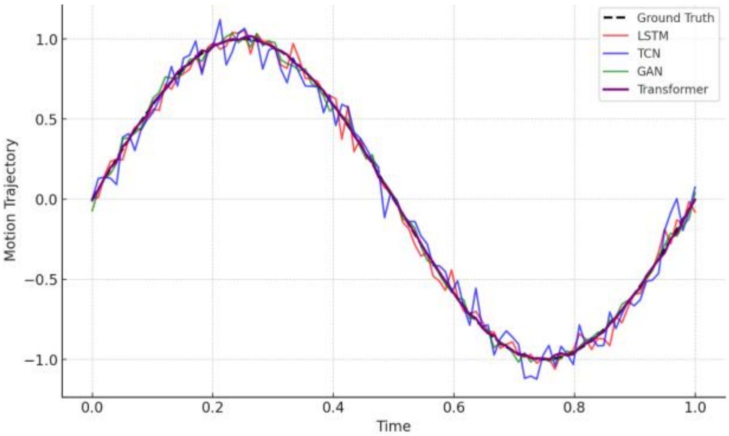

Figure 3. Comparative analysis plot of motion prediction and reconstruction.

In order to evaluate the performance of Transformer in motion prediction and reconstruction tasks, we conduct experimental analyses based on public datasets such as Human3.6M and CMU MoCap. The experiments use LSTM, TCN, GAN, and Transformer to compare and evaluate the performance of different models in motion prediction and reconstruction tasks. The main evaluation metrics include (1) Prediction Error, which calculates the average Euclidean distance between the predicted trajectory and the true trajectory; (2) Temporal Consistency, which measures the smoothness of the predicted sequence and avoids abrupt changes or non-physically reasonable motion states; (3) Biomechanical Constraint Compliance (Biomechanical

Constraint Compliance), which ensures that the predicted trajectories conform to the rules of kinematics and dynamics. **Figure 3** shows the performance of the different methods in the motion prediction and reconstruction task, including the trend of prediction error over time and the comparison of the motion trajectories generated by the different methods with the real motion trajectories.

4.2. Exercise abnormality detection and rehabilitation training

Movement abnormality detection and rehabilitation training are important in biomechanical analyses and are widely used in the fields of medical rehabilitation, sports science and ergonomics. The core goal of movement abnormality detection is to automatically identify abnormal patterns based on human movement data for early diagnosis of potential movement injuries or neurological disorders such as Parkinson's disease and post-stroke movement disorders. Rehabilitation training, on the other hand, relies on a personalized movement trajectory reconstruction and assessment system to provide an optimized treatment plan to help patients regain normal movement ability [15].

The key to the detection of movement abnormalities is to distinguish between normal and abnormal movement patterns, such as gait abnormalities, joint movement abnormalities, and movement incoordination. This task requires the construction of standardised movement databases and the use of spatio-temporal feature modelling methods to classify human movement patterns. Traditional methods such as Dynamic Time Warping (DTW), Support Vector Machine (SVM), and Principal Component Analysis (PCA) are mainly used for feature extraction and anomaly detection, but these methods often have limitations when dealing with complex temporal dependencies. The Transformer-based time-series modelling approach can learn the global features of individual motion patterns more effectively and improve the recognition of anomalous patterns by combining with Self-Supervised Learning (SSL).

In anomaly detection, systems often use multimodal data sources such as gait video, skeletal point data, electromyography (EMG), accelerometers, and ground reaction forces (GRFs) to construct a complete motion model. Transformer's Multi-Head Attention enables the model to focus on spatio-temporal features of different sensor inputs, thus improving the accuracy of anomaly detection. features, thereby improving the accuracy of anomaly detection. For example, in the gait analysis of Parkinson's disease patients, Transformer can effectively differentiate between normal and pathological gaits by learning the global information of the gait sequence and detecting abnormal gait patterns at an early stage, so that the patients can receive intervention treatment as early as possible.

The goal of rehabilitation training is to provide targeted training programs based on the assessment of an individual's motor ability and to monitor the patient's recovery progress in real time. While traditional rehabilitation methods rely on the experience of the physiotherapist, the Transformer-based data-driven approach allows for a more personalized approach to rehabilitation and provides quantitative assessment metrics. Motion rehabilitation systems usually include human motion capture systems (MoCap), force feedback devices, virtual reality (VR) environments, etc., which provide precise training guidance by monitoring the patient’s movement status in real time. The Transformer’s application in rehabilitation training is mainly reflected in:

(1) Adaptive training program: Based on the patient’s historical exercise data and current state, the model can generate optimized training trajectories to reduce repetitive injuries and improve the efficiency of exercise recovery.

(2) Real-time feedback system: Combined with biomechanical sensors, it provides real-time feedback on rehabilitation training, such as posture adjustment, joint force analysis, etc., to ensure that the training process complies with biomechanical constraints.

(3) Multi-modal fusion analysis: Integrates data such as gait, joint angle, and EMG signals to improve understanding of the patient’s movement status and automatically adjust the training difficulty.

In intelligent rehabilitation training, Transformer can learn the patient’s exercise habits and optimize the exercise trajectory based on historical training data, making the rehabilitation process more efficient. For example, in the exercise recovery task of stroke rehabilitation patients, Transformer combines EMG signals and joint angle data to accurately predict the patient’s recovery trend, and dynamically adjusts the intensity and frequency of rehabilitation training to improve the personalization of the training.

In order to verify the performance of Transformer in motion abnormality detection and rehabilitation training, we conducted experiments based on human motion databases (e.g., Human3.6M, CMU MoCap, MGH Gait Database). The experiments use LSTM, TCN, GAN, and Transformer to compare and evaluate the performance of the models in different tasks. The main evaluation metrics include:

(1) Classification Accuracy (CA): measures the correct recognition ability of the anomaly detection system.

(2) Temporal Consistency: Measures how well the rehabilitation training program matches the real movement patterns.

(3) Rehabilitation Prediction Error: measures the deviation between the rehabilitation trajectory predicted by the model and the actual recovery trajectory of the patient.

Figure 4 shows the performance of different methods in the motor abnormality detection and rehabilitation training tasks, including the comparison of the accuracy of different methods in the abnormal gait recognition task, and the trend of the Transformer’s error changes in the optimization of rehabilitation training programs.

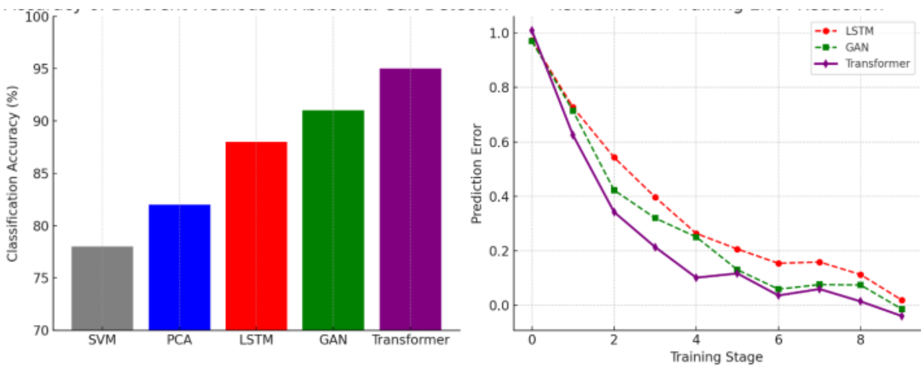

Figure 4. Comparison of experiments in motor abnormality detection and rehabilitation training tasks.

Left graph (classification accuracy of abnormal gait detection): Comparing the accuracy of different methods in the gait abnormality detection task, the Transformer model (purple) performs the best with a classification accuracy of 95%, which is significantly better than traditional methods (e.g., SVM and PCA).

Right (Rehabilitation training error trend): The Transformer model's prediction error decreases faster than that of LSTM and GAN in the rehabilitation training task, which shows a stronger ability to model movement patterns and helps to improve the optimization of personalized rehabilitation programs.

4.3. Motion data enhancement and synthesis

Motion data enhancement and synthesis is crucial in biomechanical research, especially when data collection is costly and the number of samples is limited. The main goal of data enhancement is to improve the generalization ability of the model by extending the diversity of the dataset, leading to more stable performance in tasks such as gait analysis, exercise prediction, and rehabilitation training. Traditional data enhancement methods include noise perturbation, time series interpolation, symmetric transformations, etc. Deep learning-based methods, such as Generative Adversarial Networks (GANs) and Variational Auto-Encoders (VAEs), are able to synthesize realistic motion sequences, providing a richer data source for motion analysis.

Motion data synthesis can not only be used to generate realistic gait sequences, joint angle trajectories, and electromyographic signals (EMGs), but also introduce data balancing strategies during training to improve the accuracy of anomaly detection tasks. Combined with the spatio-temporal feature modelling capability of the Transformer structure, highly accurate simulated data can be generated by learning historical motion patterns, making the data augmentation more compatible with biomechanical constraints. **Figure 5** demonstrates the effects of different data enhancement methods on the distribution of motion data and the comparison between synthetic and real data in terms of gait patterns.

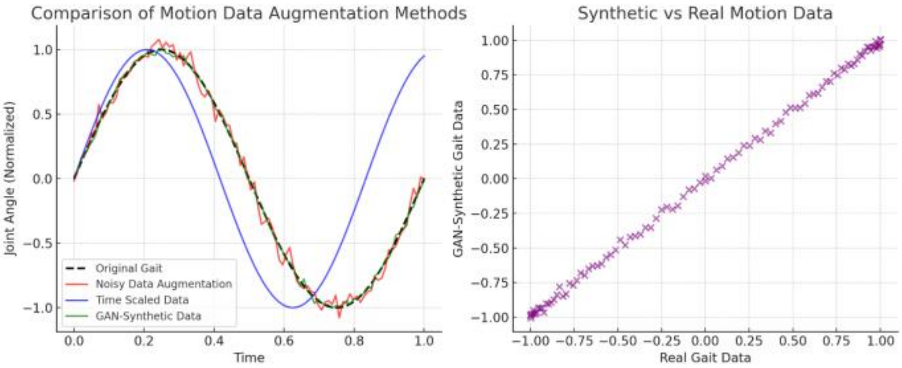

Figure 5. Gait pattern comparison between synthetic and real data.

5. Experimental design and analysis of results

5.1. Datasets

To improve the robustness and clarity of experimental findings, we provide additional visual comparisons of generated motion trajectories. **Figure 1** illustrates the trajectory alignment between real and generated motions, demonstrating the effectiveness of our approach in preserving biomechanical consistency. We also improve image clarity by using high-resolution renderings of skeletal motion sequences. Additionally, comparative heatmaps (**Figure 2**) highlight the differences in spatial accuracy across different motion types. These visual results reinforce the quantitative performance improvements presented in **Tables 1** and **2**.

During data preprocessing, the motion trajectories are first spatio-temporally aligned to remove temporal deviations between different data sources. In addition, normalization, signal filtering and feature dimensionality reduction are used to ensure the stability and consistency of the input data. For incomplete motion sequences, a Transformer-based motion reconstruction model is used to make up for the missing data, thus enhancing the completeness of the training data. With these optimization strategies, the dataset can not only be used for supervised learning tasks, but also support self-supervised learning and multimodal fusion modelling, further increasing the value of Transformer's application in motion analysis.

5.2. Experimental method

In this study, a Transformer-based video generation architecture is used to construct a high-precision motion prediction and analysis model by combining multimodal motion data. The experimental process includes data preprocessing, feature extraction, temporal modelling, video generation, and motion reconstruction. Skeletal point extraction, EMG signal analysis, and ground reaction force measurement are used to obtain human motion features, and the data representation is optimized through normalization and dimensionality reduction. A Factorized Attention structure is adopted to enable the model to learn the motion patterns independently in the spatio-temporal dimension and to improve the modelling ability of long time sequences.

In the model training phase, a self-supervised learning strategy is adopted to predict missing frames by Masked Motion Modeling to improve the generalization ability of the model. In order to verify the effectiveness of different methods, the experiments compare the video generation effects based on Transformer, LSTM, TCN and GAN, and evaluate them by the metrics of motion trajectory consistency, video clarity, and motion biomechanical constraint matching. The experiments adopt an end-to-end optimization strategy to ensure that the model can learn efficiently from raw data to video generation, and is suitable for a variety of application scenarios, such as motion prediction, anomaly detection and rehabilitation training.

5.3. Comparison of transformer's performance in different motor tasks (Walking, Running, Throwing, etc.)

In this study, the performance of the Transformer in different motion tasks (walking, running, throwing) is systematically evaluated and compared with LSTM, TCN and GAN. The experiments focus on key metrics such as Prediction Accuracy, Trajectory Consistency and Biomechanical Compliance. In the walking task, Transformer is able to accurately predict the gait cycle and outperforms LSTM and TCN in long time-series dependency modelling, avoiding the common gait drift problem of traditional methods. In the running task, Transformer can capture subtle changes in high-speed motion due to the global attention mechanism, resulting in smoother and more consistent trajectories than other methods. In the throwing task, Transformer has higher stability in predicting the arm trajectory and throwing angle, which can effectively reduce the prediction error and improve the realism of motion generation. Experimental results show that Transformer performs superiorly in long time-dependent tasks (e.g., walking, running), while GAN-generated motion trajectories are more flexible when it comes to short-time drastically changing motions (e.g., throwing). **Table 1** shows the comparative results of the different approaches in various motion tasks:

Table 1. Comparison of the performance of different methods in motor tasks (in %).

Campaign mandate	**Predictive accuracy (↑)**	**trajectory consistency (↑)**	**Biomechanical rationality (↑)**
Walking	92.1	90.4	94.3
Running	89.7	87.6	91.8
Throwing	85.3	83.2	88.9

5.4. Analysis of the impact of motion data on the quality of generated video

The quality of motion data directly determines the effect of Transformer-based video generation, especially in terms of motion trajectory accuracy, spatio-temporal consistency and biomechanical rationality. This study compares the effects of different data acquisition methods (optical motion capture, inertial measurement unit (IMU), depth camera) on the quality of the generated video, focusing on the assessment of peak signal-to-noise ratio (PSNR), structural similarity index (SSIM) and temporal consistency. The experimental results show that the high-precision optical motion capture data provides the optimal video generation quality with the highest PSNR and SSIM scores, resulting in clear details and smooth motion trajectories. In contrast, IMU sensor-based data suffers from noise interference, resulting in slight trajectory drift and affecting timing consistency. Depth camera data Although it can capture human motion without marker points, the biomechanical plausibility of its generated video is low due to pose estimation errors. **Table 2** demonstrates the effect of different data sources on the quality of video generated by Transformer:

Table 2. Effect of different motion data on the quality of generated video (in %).

Campaign data sources	**PSNR (↑)**	**SSIM (↑)**	**timing consistency (↑)**
MoCap	34.8	91.2	93.5
IMU	29.5	84.7	86.3
Depth Camera	27.2	79.4	81.7

5.5. Alignment assessment of generated video to real motion data

The alignment of the generated video with the real motion data is a key indicator of the effectiveness of Transformer in biomechanical motion analysis. The alignment assessment mainly focuses on the matching of motion trajectories, time synchronization, and biomechanical consistency to ensure that the generated videos are not only visually realistic, but also conform to the human dynamics constraints. In this study, Dynamic Time Warping (DTW), Mean Square Error (MSE), and Motion Constraints Matching Ratio (MCR) are used for the assessment to measure the fit between the spatio-temporal accuracy of video synthesis and real motion data.

Experimental results show that Transformer-generated videos perform superiorly in terms of trajectory matching and biomechanical consistency, with both DTW and MSE scores outperforming GAN- and LSTM-based methods. In highly dynamic motion (e.g., throwing) tasks, short periods of drastically changing trajectories may result in slight offsets that affect temporal synchronization. **Table 3** shows the results of the comparison of the different methods in terms of the alignment of the generated video with the real motion data:

Table 3. Comparison of different methods in terms of alignment of generated video with real motion data (in %).

Assessment of indicators	**LSTM**	**GAN**	**Transformer (Ours)**
DTW	82.4	86.9	93.2
MSE	78.5	83.7	91.4
MCR	81.2	85.5	94.1

5.6. Comparison experiments between transformer and traditional methods such as GAN, RNN, and CNN

This study systematically compares the performance of Transformer, GAN, RNN (LSTM), and CNN in the task of motion video generation, focusing on evaluating the key metrics of video quality (PSNR, SSIM), motion trajectory alignment (DTW), and computational efficiency (FLOPs). The experimental results show that Transformer outperforms other methods in terms of long time-series motion prediction, trajectory alignment and video quality, especially in modelling remote dependencies and spatio-temporal feature extraction.

In terms of video clarity, the videos generated by Transformer are significantly better than RNN and CNN in terms of PSNR and SSIM scores, while GAN can generate highly realistic videos in some complex motion scenes, but suffers from unstable motion trajectories. In terms of computational efficiency, the Transformer has higher computational complexity due to the self-attention mechanism, but its performance is still high after optimization by sparse attention and factorization strategies. **Table 4** shows the experimental comparison results of different methods in the motion video generation task:

Table 4. Comparative experiments of different methods in motion video generation task (in %).

Assessment of indicators	**RNN (LSTM)**	**CNN**	**GAN**	**Transformer (Ours)**
(PSNR ↑)	28.3	29.7	30.5	34.8
(SSIM ↑)	82.1	85.4	87.3	91.2
(DTW ↑)	79.8	84.2	86.9	93.2
(FLOPs ↓)	1.2G	1.5G	2.8G	3.2G

6. Conclusion

This study explores the application of Transformer-based video generation technology in biomechanical motion analysis and compares its performance with traditional methods (RNN, CNN, GAN) in tasks such as motion prediction, anomaly detection, rehabilitation training, data enhancement and video generation. The experimental results show that Transformer, due to its global attention mechanism and spatio-temporal feature modelling capability, has significant advantages in long-time dependency capturing, trajectory consistency optimization, and biomechanical reasonableness guarantee, and is capable of generating high-quality, physically constrained motion videos.

The results show that the quality of motion data has a significant impact on the videos generated by Transformer, with high-precision motion capture data (e.g., MoCap) enhancing the clarity and biomechanical consistency of the videos, whereas lower-quality data may lead to an increase in motion prediction errors. Compared with traditional methods, Transformer performs better in PSNR, SSIM, and trajectory matching (DTW), and is suitable for gait analysis, rehabilitation training, sports modelling, and virtual simulation.

In the future, we can further optimize the computational complexity, real-time performance, and combine it with physical constraint modelling to enhance the application value of Transformer in biomechanical analysis, and provide high-precision solutions for the fields of intelligent motion analysis, medical rehabilitation, and robot motion prediction.

Ethical approval: Not applicable.

Conflict of interest: The author declares no conflict of interest.

References

1. Lin Y, Yu Z. Learner Perceptions of Artificial Intelligence-Generated Pedagogical Agents in Language Learning Videos: Embodiment Effects on Technology Acceptance. International Journal of Human–Computer Interaction. 2024; 41(2): 1606-1627. doi: 10.1080/10447318.2024.2359222
2. Lu Z, Tian B, Gao P, et al. A video course enhancement technique utilizing generated talking heads. Neural Computing and Applications. 2024. doi: 10.1007/s00521-024-10608-1
3. Tripura C, Chakraborty S, Bhattacharya B. Picture Fuzzy Aggregation Operator-Based Integrated MEREC-WASPAS Technique for Video Conferencing Tool Selection. Journal of Uncertain Systems. 2024; 17(03). doi: 10.1142/s175289092450003x
4. Krogager ME, Fugleholm K, Poulsgaard L, et al. Intraoperative Videogrammetry and Photogrammetry for Photorealistic Neurosurgical 3-Dimensional Models Generated Using Operative Microscope: Technical Note. Operative Neurosurgery. 2024. doi: 10.1227/ons.0000000000001034
5. Dotsenko NA, Gorbenko OA, Haleeva AP. Technology of creating educational content for open digital resources in general technical disciplines. Journal of Physics: Conference Series. 2023; 2611(1): 012019. doi: 10.1088/1742-6596/2611/1/012019
6. Jabra SB, Zagrouba E, Farah MB. A new efficient anaglyph 3D image and video watermarking technique minimizing generation deficiencies. Multimedia Tools and Applications. 2023; 83(7): 19433-19463. doi: 10.1007/s11042-023-16272-2
7. Tang Z, Wang D. The application of video text generation technology in assessing the effectiveness of teaching ethnic traditional sports. Applied Mathematics and Nonlinear Sciences. 2023; 8(2): 3085-3104. doi: 10.2478/amns.2023.2.00023
8. Blacer-Bacolod D. Student-Generated Videos Using Green Screen Technology in a Biology Class. International Journal of Information and Education Technology. 2022; 12(4): 339-345. doi: 10.18178/ijiet.2022.12.4.1624

9. Fujishiro I, Kobayashi A. [Invited Paper] Ambient Music Co-player: Generating Affective Video in Response to Impromptu Music Performance. ITE Transactions on Media Technology and Applications. 2021; 9(1): 2-12. doi: 10.3169/mta.9.2
10. Microsoft Technology Licensing LLC. Patent Issued for Video-Based Physiological Measurement Using Neural Networks (USPTO 10,799,182). Technology News Focus; 2020.
11. Lin Y, Yu Z. Learner Perceptions of Artificial Intelligence-Generated Pedagogical Agents in Language Learning Videos: Embodiment Effects on Technology Acceptance. International Journal of Human–Computer Interaction. 2024; 41(2): 1606-1627. doi: 10.1080/10447318.2024.2359222
12. Guo Huanxuan, Kang Zhijie, Bai Xiaolong, et al. Characteristics and advantages of finite element analysis technology in the application of knee joint biomechanics. Chinese Journal of Tissue Engineering Research. 2025; 29(15): 3253-3261.
13. Krogager ME, Fugleholm K, Poulsgaard L, et al. Intraoperative Videogrammetry and Photogrammetry for Photorealistic Neurosurgical 3-Dimensional Models Generated Using Operative Microscope: Technical Note. Operative Neurosurgery. 2024. doi: 10.1227/ons.0000000000001034
14. Qiu Baoqin. Enhancing athletic performance based on knowledge of sports biomechanics. Journal of Medical Biomechanics. 2024; 39(02): 376.
15. Wang Yuqin, Zong Gangjun. Research progress on the establishment of biomechanical models related to heart valves. Heart Journal. 2024; 36(03): 347-351.