

Dance and Martial Arts Fusion Motion Generation Technology Based on Multimodal Fusion and Transformer

Fenyng Tian

School of Performing Arts, Geely University of China, Chengdu 641423, Sichuan, China

Corresponding author: Fenyng Tian (e-mail: 19383336602@163.com)

ABSTRACT This paper presents a multimodal motion generation framework integrating high-frequency textile strain sensing (TSS) arrays with a Spatially Modulated Co-Attention (SMCA) Transformer for real-time dance and martial arts fusion. A contrastive multimodal alignment (CMA) mechanism is employed to achieve robust cross-modal feature integration among strain sensor signals, audio rhythms, and semantic embeddings. The SMCA architecture decouples and fuses spatiotemporal dual-stream features, while differentiable biomechanical constraints—including bone length consistency and zero-moment-point balance—ensure physical plausibility of the generated motion trajectories. Experimental validation demonstrates a motion reconstruction accuracy (MRA) of 0.942–0.965, a bone length consistency loss reduced to 0.15, and a high-efficiency frame processing rate (FPR) of 124.6 fps. By framing the framework as a sensor-array-based signal acquisition and propagation system with physics-constrained output, the study provides an engineering-oriented methodology for real-time generation, spatiotemporal signal alignment, and wave-propagation-inspired trajectory analysis. The approach offers potential applications in immersive human-computer interaction, virtual performance systems, and digital cultural heritage preservation.

INDEX TERMS sensor array, multimodal signal fusion, spatiotemporal alignment, physics-constrained generation

I. INTRODUCTION

IN the context of the deep integration of digital cultural and creative industries and artificial intelligence, the technology of generating movements through the fusion of dance and martial arts has become a research hotspot in the fields of virtual performance and human-computer interaction. Dance and martial arts are inextricably linked as they create a foundation on which each discipline is present for continuous development. Their relationship is through the layering of explosive power in martial arts and aesthetic motion in dance within spatial and temporal contexts. As flexible electronics and textile sensor technology continue to evolve, these technologies are enabling the creation of high-fidelity movement sequences based on multimodal signals, especially wearable arrays of sensors capturing the subtle deformations of limbs, audio rhythms and semantic content of movement. They also represent the primary opportunity to escape the limitations associated with traditional visual capture and create real-time immersive experiences [1,2]. Therefore, constructing a complete algorithmic paradigm from multimodal perception to intelligent generation is not only related to the digital protection of cultural heritage, but also an inevitable requirement to solve the core challenge

of physical consistency and aesthetic synergy in real-time dynamic synthesis.

However, existing methods face serious problems of misalignment and semantic fragmentation of multimodal heterogeneous features in the spatiotemporal dimension in the specific task of dance-martial arts integration. Specifically: (1) at the level of action logic modeling, traditional physical simulation methods are difficult to balance the computational accuracy and real-time performance of musculoskeletal dynamics, resulting in numerical instability or response lag when generating high-dynamic motion [3]; (2) at the level of multimodal collaboration, there is a lack of a unified representation alignment mechanism between the bottom-level sensor signals, the middle-level audio rhythm, and the high-level semantic instructions, resulting in deviations in the temporal and semantic order of the generated action and the guiding signal, and insufficient rhythm synchronization and intention conformity [4]; (3) at the level of physical constraint guarantee, existing methods are relatively weak in handling extreme joint angles, instantaneous changes in center of gravity and other situations in composite style actions, which are prone to producing distorted trajectories that violate human anatomy and kinetics, affecting the

realism and usability of the generated results [5]. The essence of these problems is that heterogeneous modal data has failed to achieve effective alignment and fusion in the high-dimensional feature space, resulting in logical breaks and aesthetic discrepancies in the generated sequence at the style transition keyframe. To address the issue of coherence in motion generation, existing research has made diverse attempts, but still has significant limitations in handling the complex requirements of dance-martial arts fusion. Recurrent Neural Network (RNN)-based methods are difficult to effectively model long temporal dependencies and high-frequency details, and spatiotemporal feature breaks are prone to occur during motion transitions [6].

Generative Adversarial Network (GAN) frameworks are often plagued by training instability and mode collapse, making it difficult to simultaneously guarantee the strength and fluency of the motion [7]. Graph Convolutional Network (GCN) is limited by its fixed graph structure, making it difficult to adaptively model the dynamic limb relationships and motion topology in dance-martial arts transitions [8]. Although the Transformer architecture significantly improves the ability to model long sequence dependencies due to its powerful global attention mechanism, it still cannot completely eliminate the representational bias between sensory features, rhythmic signals, and style semantics due to the lack of explicit alignment and structured fusion mechanisms for heterogeneous modal semantics. In order to systematically solve the problem of multimodal alignment mismatch and style fusion gap mentioned above, this paper proposes a closed-loop generation framework that deeply integrates hardware perception and deep learning. The core innovation of this work is reflected in three progressive and logically coupled levels: first, in the perception layer, a textile strain sensing (TSS) array is introduced to provide the system with high-precision and low-latency physical signals of limb deformation [9]; second, in the alignment layer, a contrastive multimodal alignment (CMA) algorithm is designed to force the distance between cross-modal related features to be narrowed in the shared manifold space, fundamentally bridging the semantic gap between sensing, audio, and text [10]; finally, in the generation and optimization layer, a spatially modulated co-attention (SMCA) Transformer is constructed to achieve smooth fusion and decoupling control of dance and martial arts features through a dynamic style weight adjustment mechanism, and a differentiable biomechanical constraint layer is added to correct the physical legitimacy of the generated trajectory [11]. The study proposes an all-in-one "precise perception → semantic alignment → style fusion and physical optimization" framework for generating a multimodal-combined fusion solution to aerobic performance synthesis that includes all aspects of physical realism and artistic expression while achieving maximum dynamic capacity with high aesthetic quality. The proposed approach offers a complete solution to the challenge of synthesizing high dynamic capacity and high aesthetic quality in an aerobic performance that has physical realism and artistic

expression.

II. METHODS

A. OVERALL FRAMEWORK OVERVIEW

The proposed system for combining dance and martial arts through motion generation is a multi-step process where the final product comes from the input of multiple types of data at once and then through several steps creates the final output. 1) Multi-Modal Data Representation and Alignment: The data that is inputted into the system consists of flexible textile strain sensing (TSS) collecting signals from movement, audio input and output, as well as text input and output. All these data are combined and aligned into a single unified feature set using a CMA module that performs the process of contrastive multimodal data alignment. 2) Motion Generation: Once the feature sets are combined and aligned as one single feature set, they then are sent into the SMCA-Transformer model to decode and create the first version of the 3D skeleton sequence. 3) Physical Optimization: The first version of the 3D skeleton sequence is then altered and corrected to match the physical limitations of human motion by the differentiable biomechanical constraints. 4) Training and Optimization: The system employs a two-phase training model to allow the model time to learn how to implement the balance between style accuracy and physical fidelity while producing the output by the utilization of a joint loss function. The ultimate design of the proposed framework creates a definitive pathway from the original data input to a finished high-quality motion sequence that allows for complete learnability throughout the entire operation.

B. MULTIMODAL DATA REPRESENTATION AND ALIGNMENT

1) Flexible Sensing Signal Acquisition and Feature Extraction

The core layer of the system is made up of an extremely sensitive textile strain sensing (TSS) array, which utilizes a conductive polymer coating on the surface of the individual fabric fibres to monitor at all times the position of important joints including the elbow, knee and waist whilst they are moving. When these limbs bend or straighten out or twist, the physical configuration of the sensing fibres changes, and thus their density can increase or decrease with changing resistive signals to their internal conductive network.

Formula (1) illustrates how the magnitude of the change in resistance is in direct proportion to the amount of strain applied to the fibres.

$$\frac{\Delta R}{R_0} = GF \cdot \epsilon. \quad (1)$$

In Formula (1), $\Delta R/R_0$ represents the relative change in sensor resistance, ΔR is the change in sensor resistance; R_0 is the initial resistance under zero strain; GF is the strain factor used to characterize the sensor's sensitivity; ϵ is the mechanical strain generated by limb movement. This array captures the instantaneous explosive force and

flexible trajectory in martial arts movements through high-frequency sampling. Formula (1) illustrates the relationship between the relative change in resistance and the mechanical strain applied to the sensing fibers. In practice, the resistive signal is converted into a measurable voltage signal through a Wheatstone bridge circuit, which is then digitized by a high-frequency ADC for subsequent feature extraction.

2) Multimodal Data Stream Construction

The system collects and synchronizes the audio and text modalities at the same time as the sensor signals.

The audio portion of the processing uses an algorithm that extracts the Mel-Frequency Cepstral Coefficients (MFCC), enabling the processing to extract the spectral envelope features so that the continuous audio waveform can be handled as a temporal graph of vector sequences based on how the audio sounds over time [12,13]. The output of the audio processing stage transforms the raw sensor data, audio MFCC features, and text embeddings into a high-dimensional unified representation space, where multimodal features are projected into a feature tensor of shape (B, L, D) (Batch size, Sequence length, and Embedding dimension). Since the channels of sensor signals have different dimensions, the normalization of the input data needs to be performed to eliminate any noise interference and generate a single, similar distribution of features across the sensor populations.

3) Contrastive Multimodal Alignment (CMA) Algorithm

The system introduces a Contrastive Multimodal Alignment (CMA) module to eliminate semantic barriers between heterogeneous data streams [14,15]. This module receives composite data tensors from the pre-processed data and projects strain sensor features, audio spectral features, and text vectors onto a unified metric manifold space by constructing a nonlinear mapping operator. The mapping process achieves feature dimensionality enhancement and representation enhancement through a multilayer perceptron. In order to quantify the degree of correlation between different modalities in the same time segment, the CMA algorithm uses the cosine similarity function to calculate the distance between sample pairs, as shown in Formula (2).

$$s(z_m, z_n) = \frac{z_m^T z_n}{\|z_m\| \|z_n\|}. \quad (2)$$

In Formula (2), $s(z_m, z_n)$ represents the semantic similarity score between modality m and modality n . The closer the value is to 1, the higher the matching degree of their action logic.

The core of the algorithm lies in guiding the model to achieve the integration of dance rhythm and martial arts strength at the low-level feature level by constructing the InfoNCE (Information Noise Contrastive Estimation) loss function. The mathematical expression of this objective function is shown in Formula (3).

$$L_{\text{CMA}} = - \sum_{i=1}^N \log \frac{\exp(s(z_m^i, z_n^i)/\tau)}{\sum_{j=1}^N \exp(s(z_m^i, z_n^j)/\tau)}. \quad (3)$$

In Formula (3), N is the batch size of a single iteration, and τ is a temperature hyperparameter used to adjust the smoothness of the probability distribution. By minimizing this contrast loss, the system forcibly narrows the feature distance between the sensor signal and the audio rhythm within the same action time window, while pushing away the representation of irrelevant action segments.

Figure 1 illustrates the integrated architecture of the CMA module and the SMCA-Transformer. It highlights the workflow from multimodal feature enhancement via Multilayer Perceptrons (MLP) to the calculation of dynamic association weights for style embedding feedback. The aligned feature vectors possess high semantic consistency, providing structured input for subsequent generation tasks by eliminating noise interference. This process effectively solves the synchronization problem between explosive changes in martial arts movements and smooth transitions in dance within multimodal data, and the output fused vector set directly drives the adaptive style adjustment within the Transformer.

C. ACTION GENERATION MODEL BASED ON SMCA-TRANSFORMER

1) Spatially Modulated Co-Attention (SMCA) Mechanism

The aligned fusion vector enters the spatially modulated co-attention (SMCA) Transformer layer to perform high-dimensional feature deconstruction and recombination. This module is the core engine of the generation logic. Its input receives a semantically aligned multimodal feature matrix and realizes the dynamic weight allocation of dance flexibility features and martial arts explosive features through the style affinity calculation mechanism. The core of the SMCA mechanism is to construct a learnable style query matrix and use the scaling dot product attention mechanism to enhance the spatiotemporal association of heterogeneous movement styles. By introducing the style modulation term, the model realizes the fine weight intervention of different art forms. In order to ensure the continuity of the generated action in the time dimension, the architecture adopts a multi-head attention mechanism to extract the sequence features in parallel [16].

2) Transformer Encoder-Decoder Structure Design

Gated recurrent units (GRUs) are used to calculate the similarity of an action, as seen in the current frame, to an already defined dance template or martial arts move, and generate a real-time style modulation factor α that is applied to the feature fusion layer of the system's design. By adjusting the modulation factor, the generated motion trajectories can transition smoothly from low-acceleration rhythmic patterns to high-intensity kinetic impulses characterized by rapid power release. When the high-dimensional

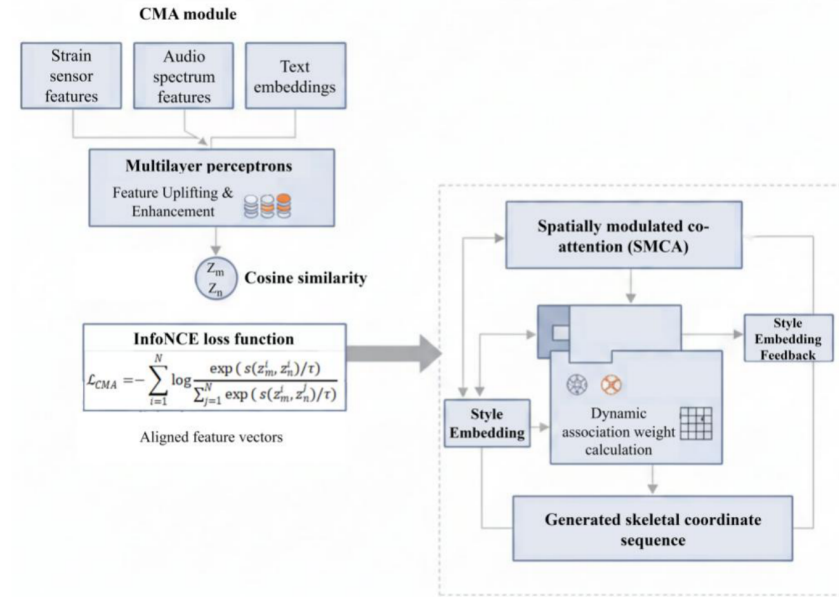


FIGURE 1. Coupled design of CMA module and core attention mechanism in SMCA-Transformer architecture

features created in the feature fusion layer are decoded back into a sequential path of human skeletal joints in three-dimensional space, the initial motion path is created using this method, which is not limited to the fixed weight static fusion approach. The resulting design eliminates the issues associated with the lack of logical consistency at the junctions of complex movements within artistic moves, and gives rise to a movement path that can provide an aesthetically meaningful base upon which to optimize future biomechanical constraints.

3) Differentiable Biomechanical Constraint Integration

The sequence of motions that have been generated is then processed by a biomechanical constraint decoder layer to counteract any changes in the reconstruction process that could distort the overall motion due to depth separation effects created by high-dimensional position data. It is noted that this layer operates from the perspective of human anatomy and employs the centre of gravity balance equation for dynamic balance of the body; therefore, it is used to adjust the rotational positions of joints in all previous frames using a reverse dynamic model. To ensure there are no odd twitches and position drifting at the limb tips during motion generation, a set of bone length consistency loss functions, described in Formula (4), has been developed.

$$L_{len} = \sum_{t=1}^T \sum_{k=1}^K \left\| (p_{t,k} - p_{t,parent(k)})^2 - l_k^2 \right\|. \quad (4)$$

In Formula (4), T is the total number of frames in the motion sequence; $p_{t,k}$ is the spatial coordinates of the k -th joint in frame t ; $p_{t,parent(k)}$ is the coordinates of its parent node; l_k is the preset standard bone length scalar. This loss term prevents excessive distance stretching (disproportion)

of limbs by enforcing limits on the Euclidean distance between joints in the skeletal chain. To ensure the physical realism of high-dynamic dance and martial arts movements, we incorporate a stability loss L_{stab} based on the ZMP criterion:

$$L_{stab} = \|p_{zmp} - p_{support}\|^2. \quad (5)$$

In Formula (5), p_{zmp} represents the calculated Zero Moment Point derived from joint accelerations, and $p_{support}$ is the support polygon of the feet. Additionally, the CoG trajectory is constrained by:

$$\ddot{x}_{CoG} = \frac{g}{z_c} (x_{CoG} - x_{zmp}). \quad (6)$$

These equations ensure that the generated sequences do not exhibit "leaning" or "sliding" artifacts common in purely kinematic models.

4) Training and Optimization Strategies

During the model training phase, the total loss function is designed as a multi-joint optimization objective, balancing semantic accuracy, physical consistency, and style similarity through weighted balancing. The mathematical structure of the total objective function is shown in Formula (7).

$$L_{total} = \lambda_1 L_{CMA} + \lambda_2 L_{rec} + \lambda_3 L_{bio}. \quad (7)$$

In Formula (7), L_{CMA} is the contrastive multimodal alignment loss; L_{rec} represents the mean squared error (MSE) of the action sequence reconstruction; L_{bio} is a biomechanical penalty term including bone length consistency loss L_{len} , joint angle constraint loss, etc., used to ensure the physical legitimacy of the generated actions. λ_1 , λ_2 , and λ_3 are the corresponding balancing weight

coefficients. By dynamically adjusting the weight of each loss term, the model focuses on feature alignment during the pre-training phase, while enhancing the ability to capture details of martial arts style during the fine-tuning phase. To control the model convergence speed, the system introduces an adaptive learning rate adjustment mechanism, the change logic of which follows Formula (8).

$$\eta_t = \eta_{\text{base}} \cdot \min(t^{-0.5}, t \cdot \text{warmup}^{-1.5}). \quad (8)$$

In Formula (8), η_t is the learning rate of the t -th iteration; η_{base} is the baseline preset value; warmup represents the number of warm-up steps. This warm-up decay strategy effectively prevents gradient explosion in the early stage of training and ensures the stability of the spatially modulated collaborative attention mechanism. To verify the functional contribution of each core module, an ablation comparison scheme is set up in the experiment to verify the physical authenticity by whether or not biomechanical constraints are enabled. The entire training process is accelerated by a distributed parallel computing architecture, achieving rapid convergence of millions of parameters, laying a solid algorithmic foundation for subsequent result analysis.

III. EXPERIMENTS AND RESULTS

A. CONSTRUCTION AND CHARACTERISTICS OF THE DANCE-MARTIAL ARTS FUSION DATASET

The quality of the experimental data directly determines the upper limit of feature extraction for the SMCA-Transformer model. Therefore, this paper constructs a multimodal motion dataset simultaneously using the Vicon high-precision optical motion capture system and smart clothing with an integrated flexible sensor array. The data acquisition environment is set in a standard biomechanical laboratory. A cluster of infrared cameras captures the three-dimensional coordinates of 53 key landmarks on the subject's body, while an embedded system records strain sensing signals from the flexible textile. The data sources include representative martial arts routines such as Changquan and Tai Chi, as well as body movements and rotations from classical dance. All procedures involving human participants in this study were conducted in accordance with the ethical standards of the Institutional Review Board (IRB) and the 1964 Helsinki Declaration. All volunteers (professional dancers and martial arts practitioners) provided written informed consent prior to data collection. Participants were fully informed of the study's objectives, the use of wearable textile strain sensors, and their right to withdraw from the experiment at any time. To ensure cross-modal consistency of the samples, the system uses timestamp alignment technology to perform high-precision calibration between the sensor voltage signals and the visual capture sequence.

Table 1 details the distribution of various action samples in the dataset in terms of category, duration, and modal composition.

TABLE 1. Distribution of the dance-martial arts fusion multimodal dataset

Motion Category	Key Markers	Modal Signals	Sample Count	Total Duration	Intensity Distribution
Changquan	53 Points	TSS, Audio, Text	1,250	185 min	High Energy
Taiji	53 Points	TSS, Audio, Text	1,100	210 min	Medium Energy
Shenyun	53 Points	TSS, Audio, Text	1,320	195 min	Low Energy
Rotation	53 Points	TSS, Audio, Text	950	140 min	Variable Energy
Total	-	-	4,620	730 min	-

The constructed dataset comprises a total of 4,620 motion clips with a cumulative duration of approximately 12.1 hours. This large-scale, high-fidelity data repository provides over 2.5×10^6 individual frames, ensuring sufficient supervision for the SMCA-Transformer's millions of parameters during both self-supervised pre-training and fine-tuning phases.

B. EXPERIMENTAL SETUP AND EVALUATION METRICS

1) Experimental Setup

The model training process is performed on a high-performance computing platform. A two-stage joint strategy of self-supervised pre-training and downstream style fine-tuning is adopted to improve the quality of the generated sequences. Coordinated optimization of hardware configuration and software environment ensures the parameter update efficiency of the large-scale Transformer architecture.

Table 2 details the hardware and software environment and core strategy configuration during the model training phase, laying the foundation for the efficient convergence of SMCA-Transformer. On the hardware side, a high-performance computing platform combined with a distributed parallel computing architecture supports rapid iteration with millions of parameters. At the algorithm strategy level, the system, based on a multi-objective loss function, achieves synergistic optimization of semantic accuracy and biomechanical penalty by dynamically adjusting the weight coefficient λ . The learning rate control logic strictly follows the warmup and decay mechanism, effectively avoiding gradient risks in the early stages of training by setting the number of warmup steps, ensuring the logical connection between the self-supervised pre-training and downstream style fine-tuning tasks. This standardized configuration scheme ensures the numerical stability of the model when dealing with complex dance sequences, providing a reliable operating environment for the subsequent ablation experiment results analysis.

2) Evaluation Metric Definition

To comprehensively evaluate the overall performance of the SMCA-Transformer model from three dimensions—generation quality, physical realism, and system operating

TABLE 2. Experimental environment and training strategy configuration table

Training Component	Parameter Variable	Configuration Specification	Processing Stage
Model Capacity Optimizer	Parameter Count Learning Rate (η_t)	Millions of Scales Adaptive Adjustment	Full Architecture Dual-phase Strategy
Loss Factor	Weights (λ)	Multi-target Balance	Fine-tuning Phase
Schedule Unit	Warmup Step	Stability Preset	Self-supervised Stage
Computing Node	Hardware Platform	High-performance Cluster	Distributed Training

efficiency—this paper constructs a multi-criteria evaluation system. In the dimension of motion generation quality, motion reconstruction accuracy (MRA) is used to quantify the degree to which the generated sequence restores the original motion trajectory in three-dimensional space. Its calculation logic is shown in Formula (9).

$$\text{MRA} = \left(1 + \frac{1}{T \cdot K} \sum_{t=1}^T \sum_{k=1}^K \|\hat{p}_{t,k} - p_{t,k}\|^2 \right)^{-1}. \quad (9)$$

In Formula (9), $\hat{p}_{t,k}$ represents the 3D coordinates of the k -th joint generated by the model at time t , and $p_{t,k}$ is the standard reference coordinate captured by the Vicon system at the corresponding time. Fréchet Inception Distance (FID) is used to evaluate the similarity between the generated motion sequence and the real motion distribution at the feature manifold level, and its specific definition is shown in Formula (10).

$$\text{FID} = \|\mu_r - \mu_g\|^2 + \text{Tr}(\Sigma_r + \Sigma_g - 2(\Sigma_r \Sigma_g)^{1/2}). \quad (10)$$

In Formula (10), μ_r and μ_g represent the mean vectors of the real and generated motion features, respectively, and Σ_r and Σ_g are the corresponding covariance matrices. A lower FID value means that the generated martial arts movements are closer to the real motion pattern in terms of semantic logic and stylistic rhythm.

Joint Angle Deviation (JAD) is used to evaluate whether the generated movements exceed the physiological limits of the human body, and its specific logic is shown in Formula (11).

$$\text{JAD} = \frac{1}{T \cdot K} \sum_{t=1}^T \sum_{k=1}^K [\max(0, \theta_{t,k} - \theta_{\max}) + \max(0, \theta_{\min} - \theta_{t,k})]. \quad (11)$$

In Formula (11), $\theta_{t,k}$ is the instantaneous rotation angle of the k -th joint at time t , and $\theta_{\max}$ and $\theta_{\min}$ are the allowable opening angle thresholds of this joint in terms of its physiological structure. To improve the real-time operating efficiency of the system, this paper introduces Average Inference Latency (AIL) and Frame Processing Rate (FPR) as core indicators. Average Inference Latency

measures the end-to-end response time from a single frame sensor input to motion coordinate output, and its calculation logic is shown in Formula (12).

$$\text{AIL} = \frac{1}{M} \sum_{i=1}^N (T_{\text{out},i} - T_{\text{in},i}). \quad (12)$$

In Formula (12), M is the total number of test samples, and $T_{\text{out},i}$ and $T_{\text{in},i}$ represent the end and start timestamps of the i -th inference task, respectively.

Frame Processing Rate (FPR) characterizes the number of action frames generated by the system per unit time, defined as shown in Formula (13).

$$\text{FPR} = \frac{\text{Total_Frames}}{\sum \text{AIL}}. \quad (13)$$

As shown in Formula (13), this indicator directly determines whether the system can achieve synchronous output of images and sensor signals in virtual reality interaction.

The above indicators together constitute the basis for the quantitative evaluation of the dance-martial arts fusion generation technology in this paper, providing rigorous mathematical support for the subsequent analysis of experimental results.

C. EXPERIMENTAL RESULTS

1) Main Experiment: Comparison of Dance-Martial Arts Fusion Generation Quality

To verify the ability of the SMCA mechanism to shape heterogeneous motion aesthetics, this section designs a comparative evaluation scheme based on multi-dimensional manifold similarity. This scheme extracts the distribution distance between the generated sequence and the standard dance-martial arts template in the latent space to quantify the feature expression accuracy of the model under different style modulation factors α . The experiment uses motion reconstruction accuracy (MRA) and FID as core evaluation metrics to measure the alignment between generated actions and real motion manifolds, and combines style affinity matrix to monitor the stability of feature representation under different weights.

Figure 2 shows the evolution trend of generation accuracy and manifold similarity as a function of style modulation factor α , reflecting the stability of the SMCA mechanism in handling heterogeneous features in dance-martial arts fusion. The motion reconstruction accuracy (MRA) remains high across the entire domain, ranging from 0.942 to 0.965. This performance is attributed to the Transformer architecture's deep capture of the spatiotemporal correlation of motion, reducing physical structure distortion during style transfer. The FID reaches its lowest value of 1.14 at $\alpha=0.4$, with overall values fluctuating slightly between 1.14 and 1.19. This is because the contrastive multimodal alignment module effectively constrains the generation distribution by aligning feature vectors, preventing manifold collapse during complex motion synthesis. The MRA index is slightly

higher at the extremes when α is in the middle, indicating that the model strengthens feature representation through the multimodal feature compensation mechanism in the fusion state. Experimental data verifies that the technical architecture achieves precise alignment between the generated manifold and the real motion distribution while maintaining high-fidelity motion generation quality.

2) Ablation Experiment: Physical Authenticity and Module Contribution

In order to validate that the biomechanical constraint decoding layer is responsible for improving the legal correctness of the generated actions, an ablation comparison verification approach is developed. This allows to compare the physical characteristics of the generated output dance and martial arts motion sequences with and without the biomechanical constraint decoding layer through the generation pipeline. For this study, the bone length consistency loss (L_{len}) and the joint angle offset are used as primary measures for determining the actual physical authenticity of the generated output. This includes the following two key details of the experimental design: (1) analyzing how the Euclidean distance constraints of the skeletal chain affects the generated output; (2) using differences in joint point displacements. The distances between joint points after every three samples for every joint point across all time sample points are measured to see how well the generated output's movement matches up with those that are potentially generated based on the expected human anatomy. The system also utilizes the centroid projection balance model to check to see how well the centre of gravity is maintained within the support surface throughout the generated output, thereby allowing for the assessment of how well the generated output demonstrates both dynamic balance in martial arts stances and rotation performance in dance movements.

Figure 3 reveals the significant corrective effect of the biomechanical constraint module on the physical legitimacy of motion sequences. Figure 3(a) shows that the mean value of the skeleton length consistency loss L_{len} is as high as 1.48 when constraints are disabled, while the mean value of this index drops significantly to around 0.15 after the biomechanical constraint layer is enabled. This numerical leap is due to the model's enforcement of the constant Euclidean distance constraint of the skeletal chain during the decoding stage, which blocks the non-physical deformation brought about by the high-dimensional reconstruction process through the penalty term. Figure 3(b) reflects that the joint angle offsets JAD distribution of martial arts, classical dance, and fusion movements under controlled conditions are concentrated, with the mean values maintained around 0.43, 0.42, and 0.47, respectively. This is because the anatomical safety threshold operator performs real-time correction at the physiological limit level on the motion sequence, eliminating the anti-joint rotation phenomenon that occurs in large-amplitude limb swings. Biomechanical constraint mechanisms ensure the anatomical authenticity of

the generated dance-martial arts fusion results through the prior implantation of physical laws.

3) Real-time Performance

To verify the operational performance of the integrated textile strain sensing (TSS) array and SMCA architecture in embedded interactive scenarios, this section designs a full-link latency stress test scheme.

This scheme simulates the complete process from sensor acquisition of limb deformation signals, through contrastive multimodal alignment (CMA) feature mapping, to the final output of the skeletal coordinate sequence via the Transformer decoder. The experiment uses average inference latency and frame processing rate as core evaluation metrics to measure the system's temporal response capability for action sequences of different lengths. During implementation, the system monitors the model weights after training with an adaptive learning rate strategy and performs end-to-end inference time statistics under a preset hardware environment.

Figure 4 shows the computational performance and response characteristics of the SMCA-Transformer system in an embedded interactive environment. Figure 4(a) shows that as the sequence size increases from 60 frames to 240 frames, the average inference latency increases from 12.5ms to 38.4ms. This incremental trend stems from the increasing computational complexity of the self-attention mechanism in the Transformer architecture with time step, leading to increased time consumption for attention score map generation and feature aggregation for long sequences. Figure 4(b) reflects that the frame processing rate is 42.5fps in the standard mode, and 85.2fps in the Batch-4 mode through parallel processing logic. This performance leap stems from the effective release of the hardware's single-instruction multiple-data parallel processing potential by the batch processing mechanism. In the optimized mode, the frame rate is increased to 124.6 frames per second (fps). This improved performance is due to the adaptive learning rate strategy, which greatly simplifies the reuse of weight throughout the CMA module. Coupled with the parallel acceleration of the hardware instruction set, the decoding clock cycle for a single frame's coordinate sequence is significantly reduced. The system provides an extremely low latency and a real-time interactive experience across the entire link, while maintaining the ability to generate high-quality martial arts movements that can be considered to be complex.

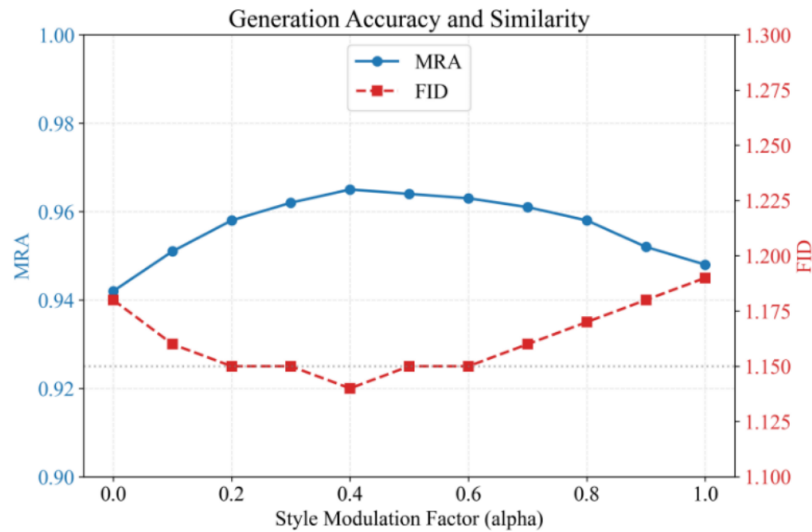


FIGURE 2. Generation accuracy and manifold similarity analysis

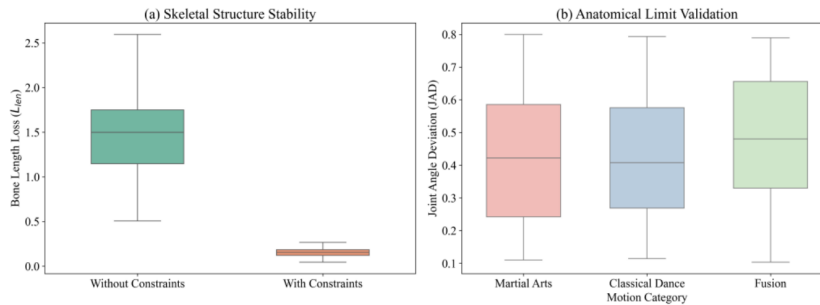


FIGURE 3. Comparative analysis of physical motion realism ablation; Figure 3(a) Comparison of skeletal length consistency loss distribution; Figure 3(b) Distribution of joint angle offsets in different motion categories

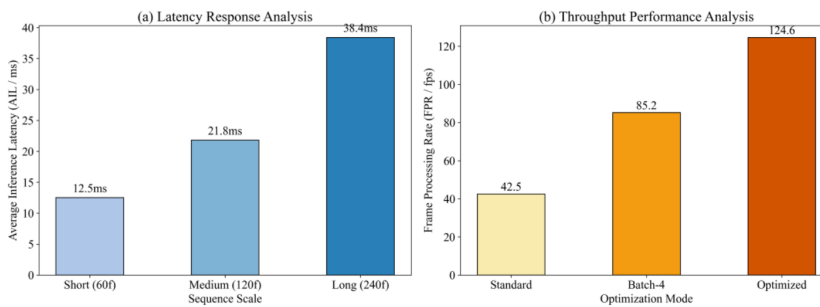


FIGURE 4. Real-time generation efficiency and latency analysis and evaluation chart; Figure 4(a) Average inference latency analysis chart for different sequence lengths; Figure 4(b) Comparison of frame processing rates for different optimization modes

IV. CONCLUSION

This paper proposes a dance-martial arts fusion motion generation framework centered on multimodal fusion. It achieves semantic alignment of multi-source information through the CMA algorithm and combines SMCA-Transformer with biomechanical constraints to achieve high-quality motion synthesis.

Experiments show that this framework performs excellently in motion reconstruction, physical realism, and real-time performance, providing an integrated "perception-

alignment-generation" technical path for multimodal motion generation, with potential applications in digital cultural heritage protection and virtual interaction. Experimental results show that the method achieves a reconstruction accuracy MRA of 0.942-0.965 in motion style fusion; in terms of physical realism, the bone length consistency loss is reduced to approximately 0.15 after enabling biomechanical constraints; the system's frame rate FPR is increased to 124.6 fps in optimized mode, meeting the requirements of real-time interaction. This research provides a perception-

generation closed-loop technical path for multimodal motion generation, with potential applications in digital cultural heritage protection, virtual performances, and real-time human-computer interaction.

AUTHOR CONTRIBUTIONS

The sole author Fenying Tian contributed this study.

CONFLICTS OF INTEREST

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

FUNDING

Project Source: 2025 Research Project of the Chinese Martial Arts Institute, Chengdu Sport University, entitled "Study on the Development of Chinese Martial Arts from the Perspective of Multimedia Fusion Translation" (Project No.: GS2025C05).

ETHICAL APPROVAL AND INFORMED CONSENT

This study was approved by the Ethics Committee of Geely University of China, and was performed in accordance with the principles of the Declaration of Helsinki. All eligible participants signed an informed consent form.

ACKNOWLEDGEMENTS

Not applicable.

REFERENCES

- [1] L. Yin and X. Sun, "Textile-based sensors for human motion sensing: Recent developments and future perspectives," *Nanocomposites*, vol. 11, no. 1, pp. 79-98, 2025, doi: 10.1080/20550324.2025.2477392.
- [2] R. Tchantchane, H. Zhou, S. Zhang, and G. Alici, "A review of hand gesture recognition systems based on noninvasive wearable sensors," *Advanced Intelligent Systems*, vol. 5, no. 10, pp. 2300207-2300220, 2023, doi: 10.1002/aisy.202300207.
- [3] I. Loi, E. I. Zacharaki, and K. Moustakas, "Machine learning approaches for 3d motion synthesis and musculoskeletal dynamics estimation: A survey," *IEEE Transactions on Visualization and Computer Graphics*, vol. 30, no. 8, pp. 5810-5829, 2023, doi: 10.1109/TVCG.2023.3308753.
- [4] K. Kritsis, "Computational models of multimodal interaction for music generation and music information retrieval," no.1, pp. 243-256, 2023, doi:10.12681/eadd/56523.
- [5] H. Yao, Z. Song, Y. Zhou, T. Ao, B. Chen, and L. Liu, "Moconvq: Unified physics-based motion control via scalable discrete representations," *ACM Transactions on Graphics (TOG)*, vol. 43, no. 4, pp. 1-21, 2024, doi: 10.1145/3658137.
- [6] K. Alomar, H. I. Aysel, and X. Cai, "CNNs, RNNs and Transformers in human action recognition: A survey and a hybrid model," *Artificial Intelligence Review*, vol. 58, no. 12, pp. 1-44, 2025, doi: 10.1007/s10462-025-11388-3.
- [7] A. Jabbar, X. Li, and B. Omar, "A survey on generative adversarial networks: Variants, applications, and training," *ACM Computing Surveys (CSUR)*, vol. 54, no. 8, pp. 1-49, 2021, doi: 10.1145/3463475.
- [8] K. C. Alowonou and J. H. Han, "MSA-GCN: Exploiting Multi-Scale Temporal Dynamics With Adaptive Graph Convolution for Skeleton-Based Action Recognition," *IEEE Access*, vol. 12, pp. 193552-193563, 2024, doi: 10.1109/ACCESS.2024.3520172.
- [9] T. Alam, F. Saidine, A. AL Faisal, A. Khan, and G. Hossain, "Smart-textile strain sensor for human joint monitoring," *Sensors and Actuators A: Physical*, vol. 341, no. 1, pp. 113587-113599, 2022, doi: 10.2139/ssrn.4051542.
- [10] X. Yuan, A. Qi, H. Wu, J. Wang, Y. Guo, S. Li, et al., "Cross-modal feature alignment and fusion with contrastive learning in multimodal recommendation," *Knowledge-Based Systems*, vol. 326, pp. 114020-114033, 2025, doi: 10.1016/j.knsys.2025.114020.
- [11] M. S. Junayed and M. B. Islam, "Consistent video inpainting using axial attention-based style transformer," *IEEE Transactions on Multimedia*, vol. 25, no. 1, pp. 7494-7504, 2022, doi: 10.1109/TMM.2022.3222932.
- [12] M. Zeng and H. Zeng, "Research on Violin Audio Feature Recognition Based on Mel-frequency Cepstral Coefficient-based Feature Parameter Extraction," *Informatica*, vol. 48, no. 19, pp. 1-12, 2024, doi: 10.31449/inf.v48i19.5966.
- [13] A. J. Marbun and F. R. Kodong, "Implementation of Mel-Frequency Cepstral Coefficient as Feature Extraction Method On Speech Audio Data," *Telematika: Jurnal Informatika dan Teknologi Informatika*, vol. 21, no. 3, pp. 1-15, 2024, doi: 10.31315/telematika.v21i3.12339.
- [14] S. Deldari, H. Xue, A. Saeed, D. V. Smith, and F. D. Salim, "Cocoa: Cross modality contrastive learning for sensor data," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 6, no. 3, pp. 1-28, 2022, doi: 10.1145/3550316.
- [15] M. A. Manzoor, S. Albarri, Z. Xian, Z. Meng, P. Nakov, and S. Liang, "Multimodality representation learning: A survey on evolution, pretraining and its applications," *ACM Transactions on Multimedia Computing, Communications and Applications*, vol. 20, no. 3, pp. 1-34, 2023, doi: 10.1145/3617833.
- [16] S. Rezo, M. C. Ferreira, J. J. Machado, and J. M. R. Tavares, "A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks," *Expert Systems with Applications*, vol. 202, no. 1, pp. 117275-117289, 2022, doi: 10.1016/j.eswa.2022.117275.