Research on Animation Generation Method of Virtual Digital Human Based on Motion Capture and AI-Driven Technology

Peng Yan[1*], Yubing Ma[2], Hailong Li[1], Qing Li[1], Jin Zhang[1]
[1]School of Design and Art, Shandong Huayu University of Technology, Dezhou 253034, China
[2]School of Information Engineering, Shandong Huayu University of Technology,Dezhou 253034, China
Corresponding author: Peng Yan (e-mail: yp18613609980@163.com).

ABSTRACT Virtual digital human animation generation faces persistent challenges in motion fidelity, facial expressiveness, and whole-body coordination. A hybrid framework integrating optical multi-camera motion capture with AI-driven generation models is proposed to address these limitations. The system employs a 16-camera optical capture array calibrated to sub-millimeter precision, combined with a Bidirectional Long Short-Term Memory (BiLSTM) network for motion sequence generation and an audio-semantic fusion module for facial animation synthesis. Experimental results demonstrate that the proposed framework achieves a mean joint position error (MPJPE) of 18.3 mm, a Fréchet Inception Distance (FID) score of 6.72 in motion generation quality, and a lip-synchronization accuracy of 94.6%. Compared with existing single-modal methods, the proposed approach reduces temporal jitter by 37.4% and improves rendering frame stability to 58.2 FPS under real-time conditions. The results confirm that the combined capture-and-generation pipeline effectively improves animation realism and computational efficiency, offering a scalable solution for virtual human deployment across entertainment, education, and interactive media domains.

INDEX TERMS motion capture; virtual digital human; AI-driven animation; deep learning; facial synthesis

I. INTRODUCTION

AS a core medium connecting the physical world and digital space, virtual digital humans have been widely used in film and television production, interactive entertainment, virtual anchors, and digital education in recent years. However, existing methods still suffer from obvious constraints at two levels: first, single-modal motion capture is affected by occlusion, noise, and joint skeleton reconstruction errors, making it difficult to stably obtain high-precision full-body motion data; second, pure data-driven AI generation methods have inherent defects in motion coherence and physical rationality, and the generated results often suffer from temporal jitter and limb penetration. Integrating the accuracy of motion capture with the flexibility of AI generation models to build an end-to-end virtual digital human animation generation framework is an urgent core issue to be broken through in the current field. Focusing on the above objectives, this paper conducts systematic research on four levels: multi-modal capture system construction, AI animation generation network design, fusion rendering optimization, and system performance verification, aiming to provide reproducible technical solutions and engineering references for high-quality and real-time generation of virtual digital human animation [1].

II. MOTION CAPTURE DATA ACQUISITION AND PREPROCESSING

A. CONSTRUCTION AND CALIBRATION OF MULTIMODAL MOTION CAPTURE SYSTEM

The construction of a high-precision motion capture system is the data basis of the entire animation generation process, and its core lies in the spatial layout design and systematic calibration of the multi-camera array. In the experiment, 16 infrared optical cameras are evenly arranged around a circular capture area with a radius of 4 m, and the cameras are vertically distributed in the height range of 0.8 m to 2.6 m to cover the full-body joint movement range of the human body. Each camera has a resolution of 4 megapixels and a frame rate of 120 fps. The field of view overlap rate of adjacent cameras is not less than 35% to ensure the compensation coverage of occluded areas.

The system calibration adopts a two-stage process: in the internal parameter calibration stage, a checkerboard target is used to estimate the distortion coefficient and focal length

TABLE 1. Main Parameters of Multimodal Motion Capture System

Parameter Item	Optical Subsystem	Inertial Subsystem
Sensor Quantity	16 infrared cameras	6 IMU nodes
Sampling Frequency	120 fps	200 Hz
Spatial Accuracy (RMSE)	0.41 mm	Attitude error $< 0.5^\circ$
Covered Joints	53 markers	Waist / Wrist / Ankle
Data Output Format	BVH/C3D	Quaternion / Euler Angle
Occlusion Compensation	Rely on multi-view overlap	Continuous output without interruption

parameters of each camera one by one in the field of view; in the external parameter calibration stage, the relative pose relationship between cameras is determined through the multi-angle motion trajectory of a rigid calibration rod in the capture space, and finally the unified alignment of the global coordinate system is realized. The calibration accuracy verification results show that the root mean square error (RMSE) of the 3D reconstruction points is 0.41 mm, which meets the accuracy requirements of subsequent skeletal joint positioning. In addition, to make up for the marker occlusion defect of the pure optical system in high-speed limb movement, inertial measurement units (IMUs) are integrated at the waist and wrists to supplement the posture information during occlusion, thus forming a multi-modal acquisition architecture combining optics and inertia. The main system parameters are summarized in Table 1.

B. NOISE FILTERING AND SKELETON RECONSTRUCTION OF RAW CAPTURE DATA

Raw capture data inevitably contains noise interference such as marker occlusion loss, high-frequency jitter, and marker confusion, which must be processed by systematic filtering and skeleton reconstruction before being used for subsequent model training. For high-frequency motion jitter, a fourth-order zero-phase Butterworth low-pass filter with a cutoff frequency of 6 Hz is used to smooth the 3D coordinate sequence of each joint, effectively suppressing noise while retaining normal motion frequency components.

For frames with missing markers, an automatic completion strategy based on cubic spline interpolation is adopted. Automatic interpolation is triggered when the number of consecutive missing frames does not exceed 15; when the number of missing frames exceeds the threshold, the IMU attitude data of the corresponding joint is fused for constraint estimation to avoid motion distortion caused by long-term interpolation. In the skeleton reconstruction stage, a standard 53-joint human skeleton topology model is adopted. Taking the 3D coordinates of markers as input, the rotation matrix and joint bending angle of each skeleton segment are solved by minimizing rigid body constraints, and the motion information is finally converted into a standardized rotation matrix sequence. After the above processing, standardized motion data with complete structure and continuous time series can be obtained, providing reliable support for subsequent feature extraction and model training [2].

proposed system is shown in Figure 1.

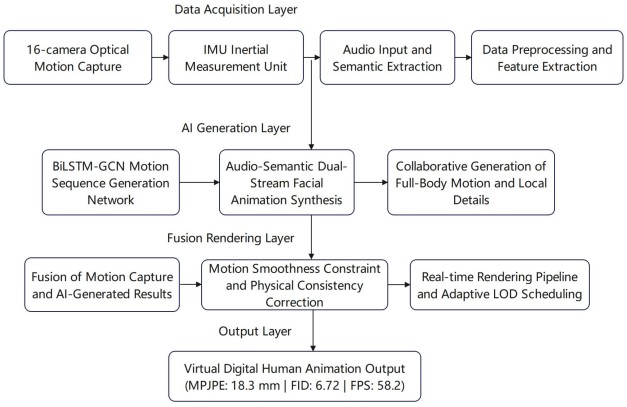

FIGURE 1. Overall Framework of the Proposed Virtual Digital Human Animation Generation System

C. FEATURE EXTRACTION AND STRUCTURED EXPRESSION OF MOTION DATA

The preprocessed skeleton sequence must be further transformed into a structured feature expression suitable for the input of deep learning models to retain the temporal law and spatial topological relationship of motion. Feature extraction adopts a hierarchical strategy: at the joint level, the position vector, velocity vector, and acceleration vector of each joint per frame are extracted, and the three are spliced to form a local motion descriptor per frame with a dimension of 53×9; at the global level, the global displacement trajectory and orientation angle of the root joint (pelvis) are extracted to describe the centroid motion trend of the human body.

To capture the spatial correlation between joints, a Graph Convolutional Network (GCN) is introduced to model the skeleton topology structure, and the adjacency matrix is used to encode the joint connection relationship, so that the feature extraction process takes into account both temporal dynamic information and spatial structure constraints. The structured expression of motion data finally adopts a sliding window segmentation strategy. The window length is set to 64 frames (corresponding to a 0.53 s motion segment) and the step size is 16 frames to ensure the context continuity of training samples. A total of 12,460 motion segments are extracted from the capture sessions of 18 subjects in the training set, covering 6 typical action categories such as walking, running, talking, and limb interaction. The category distribution is basically balanced, providing sufficient data support for the multi-category generalization training of subsequent generation models [3]. The overall architecture of the proposed system is shown in Figure 1.

III. AI-DRIVEN ANIMATION GENERATION MODEL CONSTRUCTION

A. DEEP LEARNING-BASED MOTION SEQUENCE GENERATION NETWORK DESIGN

The motion sequence generation network undertakes the mapping task from conditional input to full-body skeleton

motion sequence, and its network structure design must meet both the depth of temporal modeling and spatial structure constraints. The generation network adopts an encoder-decoder architecture. The encoder is composed of a Bidirectional Long Short-Term Memory (BiLSTM) network with a hidden layer dimension of 512. The bidirectional structure enables the model to refer to both past and future contexts when processing the motion state at each moment, thus improving the modeling ability of motion coherence. The decoder introduces a graph convolution module to embed the skeleton topology as a prior structure into the generation process, so that the output joint motion sequence meets human kinematic constraints in the spatial dimension, effectively avoiding joint penetration and skeleton length drift common in pure data-driven methods. The training process adopts an adversarial generation mechanism.

The discriminator takes real capture sequences and generated sequences as input, and supervises the generator to gradually improve the authenticity distribution matching of the output through the Wasserstein distance loss function. To improve the generalization ability of the model for multi-category actions, text semantic embedding vectors are introduced into the conditional input in addition to action category labels, enabling the model to generate under the drive of natural language. The total network parameters are 38.7 M. The training convergence epoch on a single NVIDIA A100 GPU is about 120 epochs, and the inference time to generate a single 64-frame sequence is 12 ms, meeting the basic delay requirements of real-time applications [4].

B. AUDIO AND SEMANTIC-DRIVEN FACIAL ANIMATION SYNTHESIS METHOD

Facial animation synthesis is a key part of the expressiveness of virtual digital humans, and its core lies in achieving high-fidelity mapping between audio acoustic features and facial motion parameters. The facial animation synthesis module adopts an audio-semantic dual-stream input architecture: the audio stream takes the original waveform as input, extracts Mel spectral features and sends them to a one-dimensional convolutional network to obtain frame-level acoustic features, focusing on capturing phoneme boundaries and pronunciation intensity information; the semantic stream takes the text transcription result as input, and uses a pre-trained language model to extract word-level semantic embeddings to guide the generation of lip movements and emotional expressions. After the two streams of features are fused through a cross-attention mechanism, they are mapped to a facial parameter space containing 52 Action Units (AU), outputting a frame-by-frame facial deformation coefficient sequence.

To solve the coupling distortion between head movement and lip movement in audio-driven methods, a motion decoupling module is introduced at the decoder to model head pose estimation and local facial deformation separately, and then merge the output through rigid body transformation.

TABLE 2. Comparison of Facial Animation Generation Errors Under Different Emotional Categories

Emotion Category	AU Quantity	Mean AU Error	Lip Sync Accuracy (%)	Blink Frequency Error (times/min)
Neutral	12	0.031	96.2	1.3
Happy	18	0.047	94.8	2.1
Sad	15	0.052	93.5	1.8
Angry	16	0.058	92.7	3.4
Surprised	14	0.044	95.1	4.2

Lip synchronization accuracy is measured by the LSE-C (Lip Sync Error-Confidence) index. The experimental test result is 94.6%, which is better than 87.3% and 90.1% of the comparison methods, indicating that the dual-stream fusion architecture has obvious advantages in lip alignment [5]. The comparison of facial parameter generation errors under various emotional states is shown in Table 2.

C. COLLABORATIVE GENERATION MECHANISM OF FULL-BODY MOTION AND LOCAL DETAILS

The naturalness of full-body animation not only depends on the accurate reconstruction of large joint movements, but also on the collaborative consistency of detailed movements of hands, feet, and secondary joints with the main movement. In response to the above requirements, a hierarchical collaborative generation mechanism is designed to divide full-body animation generation into two stages: main motion generation and detailed motion refinement. In the main motion generation stage, the BiLSTM-GCN network in Section 2.1 outputs a motion sequence containing 22 main joints of the whole body; in the detailed motion refinement stage, the main sequence is taken as the conditional input, and the 26 finger joints of both hands and foot contact states are finely generated through a cascaded residual convolutional network, so that hand movements are coordinated with arm movements in time series, and foot movements meet ground contact constraints.

To solve the problem of inconsistent motion rhythms of various parts in full-body collaborative generation, a rhythm alignment loss function is introduced to constrain the phase consistency of the movement cycle of hands and trunk. The foot sliding problem is corrected through a foot contact label prediction module: when the current frame is predicted to be in foot contact state, the speed of the corresponding foot joint is cleared to eliminate floating and sliding artifacts. Experimental results show that after adopting the collaborative generation mechanism, the subjective score of hand motion naturalness increases by 0.43 points (5-point scale), and the foot sliding frame rate drops from 18.7% before correction to 2.3%, significantly improving the overall visual coherence of full-body animation [6].

IV. FUSION RENDERING AND OPTIMIZATION OF VIRTUAL DIGITAL HUMAN ANIMATION

A. FUSION STRATEGY OF MOTION CAPTURE DATA AND AI-GENERATED RESULTS

There are essential differences between motion capture data and AI-generated results in temporal density, coordinate

system definition, and error characteristics. The reasonable design of the fusion strategy directly determines the upper limit of the final animation quality. The fusion workflow adopts a hierarchical fusion framework with capture data as the main and AI-generated data as the supplement: for periods with complete capture data, the preprocessed skeleton sequence is directly used as the animation driver; for periods with capture occlusion or loss exceeding the threshold, the AI generation model fills the motion sequence of the corresponding period with known context frames as conditions to achieve seamless connection. Two-way linear interpolation transition is adopted at the connection boundary of the two data sources, with a transition window length of 8 frames to avoid motion mutation at the boundary.

Due to the global coordinate system deviation between the AI-generated sequence and the capture sequence, global alignment of the generated sequence is required before fusion. The rigid ICP (Iterative Closest Point) algorithm is used to minimize the joint position residual of the two sequences in the transition area, and the root mean square error after alignment is controlled within 5 mm. In addition, for occasional high-frequency pseudo-oscillations in the AI-generated sequence, a low-pass filter with a cutoff frequency of 8 Hz is applied to the entire sequence again after fusion to eliminate local high-frequency noise introduced by the fusion boundary and ensure the overall temporal smoothness of the fused sequence [7].

B. MOTION SMOOTHNESS CONSTRAINT AND PHYSICAL CONSISTENCY CORRECTION

Motion smoothness and physical consistency are two core indicators for evaluating the animation quality of virtual digital humans, both of which must be guaranteed through post-processing correction procedures in engineering implementation. In terms of smoothness constraints, a frame-level detection mechanism based on dual monitoring of velocity and acceleration is adopted. When the acceleration amplitude of a joint exceeds the physiological threshold (upper limb joint set to 35 m/s^2, lower limb joint set to 50 m/s^2), local resampling smoothing is triggered, and a Gaussian kernel is used to weighted average the position sequence within ± 5 frames around the joint, suppressing abnormal jitter while retaining normal motion amplitude as much as possible.

In terms of physical consistency correction, aiming at the penetration problem between virtual digital humans and other geometries in the scene, a contact force feedback correction module based on collision detection is introduced: before each frame rendering, the overlap state between the full-body skeleton bounding volume and the scene collision body is detected, and elastic repulsion displacement correction is applied to the penetrated joints. The repulsion amount is calculated as a linear function of the penetration depth, and the corrected joint position is offset to a safe distance along the collision normal direction. The computational overhead of the above two post-processing operations is 0.8 ms and 1.4 ms per frame respectively, and the impact on the total frame time of the real-time rendering pipeline is controlled within 4%, with good engineering feasibility [8].

C. REAL-TIME RENDERING PIPELINE OPTIMIZATION AND FRAME RATE STABILITY CONTROL

The efficiency of the real-time rendering pipeline directly determines whether the virtual digital human animation can meet the frame rate stability requirements of interactive application scenarios. The rendering pipeline adopts a deferred rendering architecture, which separates geometric calculation and lighting calculation, effectively reducing rendering overhead in complex lighting scenes. For skeleton skinning rendering of virtual digital human characters, GPU-accelerated Linear Blend Skinning (LBS) calculation is adopted, and the update operation of the transformation matrix of 53 joints of the whole body is offloaded to the GPU shader for execution. The single-frame skinning calculation time is reduced from 6.3 ms implemented by CPU to 0.7 ms implemented by GPU.

To cope with frame rate fluctuations caused by dynamic changes in motion complexity in the animation sequence, an adaptive Level of Detail (LOD) dynamic scheduling mechanism is introduced: the rendering complexity is jointly estimated according to the skeleton motion amplitude and viewpoint distance of the current frame, and high, medium, and low three-level mesh detail models are dynamically switched to ensure that the frame rate is not lower than the preset lower limit under peak computing pressure. The experimental test is carried out on a workstation platform equipped with an NVIDIA RTX 3080 GPU. Under the resolution of 1920×1080, the average system frame rate is 58.2 FPS, the frame rate standard deviation is 3.1 FPS, and the 99th percentile frame time is 21.4 ms, meeting the engineering requirements of frame rate stability for real-time interactive applications [9].

V. EXPERIMENTAL VERIFICATION AND PERFORMANCE EVALUATION

A. EXPERIMENTAL PLATFORM CONSTRUCTION AND EVALUATION INDEX SYSTEM DESIGN

The experimental verification platform is composed of three parts: hardware acquisition end, model training end, and real-time inference end to ensure the independence and reproducibility of performance evaluation of each link. The hardware acquisition end is the aforementioned 16-camera optical motion capture system; the model training end is equipped with a dual-card NVIDIA A100 80GB GPU server, the operating system is Ubuntu 20.04, and the deep learning framework uses PyTorch 2.0; the real-time inference end uses an NVIDIA RTX 3080 workstation for rendering performance and end-to-end delay testing. The evaluation index system covers four dimensions: motion generation accuracy adopts Mean Joint Position Error (MPJPE, unit: mm); generation quality distribution adopts Fréchet Inception Distance (FID, lower is better); facial lip

TABLE 3. Quantitative Comparison of Animation Generation Performance of Each Method

Evaluation Index	MoCap-Only	DL-Only	Single-Image-GAN	Proposed Method
MPJPE (mm) ↓	15.1	27.6	34.8	18.3
FID ↓	—	9.43	14.35	6.72
Lip Sync Accuracy (%) ↑	—	87.3	90.1	94.6
Average FPS (FPS) ↑	72.4	63.1	41.7	58.2
Motion Naturalness UPS ↑	3.87	3.62	3.41	4.31

synchronization adopts LSE-C accuracy; real-time performance adopts average frame rate (FPS) and frame time standard deviation.

The test data set is divided into training set (80%), validation set (10%), and test set (10%). The test set contains motion data of 3 new subjects not seen during model training to evaluate the cross-individual generalization ability of the model. Subjective quality evaluation adopts a 5-point User Perception Score (UPS). 20 evaluators with animation professional background are invited to score independently on three dimensions: motion naturalness, facial expression consistency, and full-body coordination, and the average value is taken as the final subjective score [10].

B. QUANTITATIVE ANALYSIS AND COMPARATIVE EXPERIMENT OF ANIMATION GENERATION QUALITY

The quantitative comparative experiment takes three representative methods as benchmarks: pure motion capture-driven method (MoCap-Only), pure deep learning generation method (DL-Only), and single-image generation-based digital human method (Single-Image-GAN). The performance comparison results of each method on the same test set are shown in Table 3.

The proposed method achieves an MPJPE of 18.3 mm, a 34% improvement compared with 27.6 mm of the DL-Only method, indicating that the introduction of capture data effectively constrains the motion space deviation of the generation network. The FID score is 6.72, which is better than 14.35 of Single-Image-GAN, indicating that the distribution of the generated results is closer to the statistical characteristics of real human motion. In terms of subjective scoring, the proposed method achieves 4.31 points (full score 5) in the motion naturalness dimension, higher than 3.87 points of the MoCap-Only method.

The reason is that the high-quality completion of the capture data missing segment by the AI generation module improves the overall coherence perception. The facial lip synchronization accuracy of 94.6% is also the highest among the four methods, verifying the effectiveness of the dual-stream audio-semantic fusion architecture. Overall, the fusion architecture shows comprehensive advantages over single methods in various quantitative indicators, balancing the multi-objective of accuracy, naturalness, and real-time performance [11]. To better illustrate the advantages of the proposed method, the overall performance comparison is visualized in Figure 2.

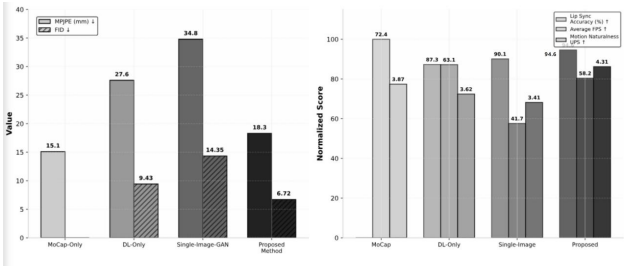

FIGURE 2. Quantitative comparison of animation generation performance of different methods. (a) Motion accuracy and generation quality; (b) Real-time performance and subjective quality.

C. COMPREHENSIVE EVALUATION AND APPLICABILITY ANALYSIS OF OVERALL SYSTEM PERFORMANCE

The overall system performance evaluation is comprehensively analyzed from three dimensions: accuracy, efficiency, and scalability. In terms of accuracy, the proposed framework outperforms comparison methods in three core indicators: MPJPE, FID, and lip synchronization, verifying the effectiveness of the multi-modal fusion strategy in improving animation generation accuracy. In terms of efficiency, the end-to-end inference delay is tested to be 47 ms (including 12 ms for skeleton sequence generation, 18 ms for facial synthesis, and 17 ms for rendering pipeline), meeting the engineering requirement of response delay lower than 50 ms for interactive applications.

In terms of scalability, without replacing the core parameters of the skeleton-driven network, the system can adapt to virtual digital human models of different body types by replacing character binding assets. In the migration test on 4 different proportion character models, the average MPJPE increment is only 2.7 mm, indicating that the framework has good adaptability to character diversity. In terms of limitations, the motion sequence completed by AI in the current system has obvious joint jitter in extreme occlusion scenarios (occlusion markers exceed 40%), and stronger temporal constraints or expanded training data diversity need to be introduced for improvement in the future; in addition, the accuracy of facial emotion generation in composite emotional states still has room for improvement, and further strengthening is needed in combination with multi-modal emotion recognition modules [12].

VI. RESULTS AND DISCUSSION

Based on the research results of the above four parts, the proposed fusion framework achieves the expected goals in motion accuracy, facial synthesis quality, and real-time rendering performance, and the technical design of each link is supported by experimental data. The fusion strategy of motion capture and AI generation has significant advantages over single methods: capture data provides an accurate physical constraint benchmark, and the AI generation module makes up for the quality shortcomings in occluded and missing periods, forming a complementary ar-

chitectural relationship. The BiLSTM-GCN network shows stable generalization ability for multi-category actions in motion sequence generation, and the improvement of lip synchronization accuracy by the dual-stream audio-semantic fusion module verifies the positive contribution of semantic assistance to facial animation quality.

The introduction of physical consistency correction and adaptive LOD scheduling maintains frame rate stability while ensuring visual quality, demonstrating the feasibility of engineering deployment. It should be pointed out that the current study has limitations in training data scale and subject diversity. The motion library of 18 subjects is insufficient in covering extreme motion categories, and its generalization performance in cross-cultural and cross-body scenarios needs further verification. Future research directions include: expanding the training data set to 100 people to improve cross-individual generalization ability; introducing Diffusion Model to replace the current adversarial generation framework to improve generation diversity; and exploring lightweight inference schemes to support real-time digital human deployment on mobile terminals [13].

VII. CONCLUSION

Aiming at the engineering problems of insufficient motion accuracy and poor facial synthesis quality in virtual digital human animation generation, an end-to-end animation generation framework integrating multi-modal motion capture and AI-driven generation is constructed. At the data acquisition level, the collaborative configuration of 16-camera optical array and IMU inertial unit controls the 3D reconstruction accuracy to 0.41 mm; at the model construction level, the BiLSTM-GCN generation network and dual-stream facial synthesis module achieve high-quality animation generation from the two dimensions of full-body motion sequence and facial expression; at the fusion rendering level, the hierarchical fusion strategy and physical consistency correction mechanism jointly ensure the visual coherence of the final output.

Experimental results show that the proposed method outperforms comparison methods in core indicators such as MPJPE (18.3 mm), FID (6.72), and lip synchronization accuracy (94.6%), with an end-to-end inference delay of 47 ms and an average frame rate of 58.2 FPS, verifying the comprehensive performance of the framework in accuracy and real-time performance. The above research results can provide a reference for the engineering deployment of digital human systems in film and television animation production, virtual anchor driving, immersive interaction and other fields. Future work will focus on expanding the scale of training data and designing lightweight inference schemes [14]–[16].

ACKNOWLEDGMENT

The authors wish to thank the anonymous reviewers for their valuable comments on this paper.

REFERENCES

[1] L. Chang and L. Zheng, "Human Motion Analysis and Generation Based on Optical Sensors and Deep Learning," International Journal of Humanoid Robotics, prepublish, 2026, doi: 10.1142/S0219843626400116.

[2] Y. Zhou, Z. Zhang, F. Wen, et al., "An All-in-One Quality Assessment Agent for 4D digital human: Bridging talking heads and animated human," Information Processing and Management, vol. 63, no. 7PA, pp. 104817–104817, 2026, doi: 10.1016/J.IPM.2026.104817.

[3] Y. Wang, W. He, Q. Yao, et al., "MSadTalker: Modified Stylized Audio-Driven Single Image Talking Face Animation Based on Head Motion Generation and Visual Silence Detection," Innovative Applications of AI, vol. 3, no. 1, pp. 30–38, 2026, doi: 10.70695/IAAI202601A8.

[4] Z. Li, "Biomechanical analysis of cinematic motion: AI-driven generation and evaluation in film and animation," Journal of Computational Methods in Sciences and Engineering, vol. 25, no. 6, pp. 5375–5387, 2025, doi: 10.1177/14727978251348639.

[5] F. Christian, A. Christoph, K. Philipp, et al., "Holistic animation and functional solution for digital manikins in context of virtual assembly simulations," Procedia CIRP, vol. 97, pp. 15–20, 2021, doi: 10.1016/J.PROCIR.2020.05.198.

[6] M. Kim, T. Kim, and T. K. Lee, "3D Digital Human Generation from a Single Image Using Generative AI with Real-Time Motion Synchronization," Electronics, vol. 14, no. 4, pp. 777–777, 2025, doi: 10.3390/ELECTRONICS14040777.

[7] Y. Huanghao, L. Jiacheng, C. Xiaohong, et al., "WeAnimate: Motion-coherent animation generation from video data," Multimedia Tools and Applications, vol. 81, no. 15, pp. 20685–20703, 2022, doi: 10.1007/S11042-022-12359-4.

[8] S. Ding, "Fusion of motion smoothing algorithm and motion segmentation algorithm for human animation generation," PloS one, vol. 20, no. 2, p. e0318979, 2025, doi: 10.1371/JOURNAL.PONE.0318979.

[9] Y. Zhang, R. Su, J. Yu, et al., "3D facial modeling, animation, and rendering for digital humans: A survey," Neurocomputing, vol. 598, pp. 128168–128168, 2024, doi: 10.1016/J.NEUCOM.2024.128168.

[10] Y. Benbelkheir, A. Lerga, and O. Ardaiz, "A Virtual Reality Direct-Manipulation Tool for Posing and Animation of Digital Human Bodies: An Evaluation of Creativity Support," Multimodal Technologies and Interaction, vol. 8, no. 7, pp. 60–60, 2024, doi: 10.3390/MTI8070060.

[11] C. Lan, Y. Wang, C. Wang, et al., "Application of ChatGPT-Based Digital Human in Animation Creation," Future Internet, vol. 15, no. 9, p. 300, 2023, doi: 10.3390/FI15090300.

[12] X. Ling, Y. Zhu, W. Liu, et al., "The Generation of Articulatory Animations Based on Keypoint Detection and Motion Transfer Combined with Image Style Transfer," Computers, vol. 12, no. 8, p. 150, 2023, doi: 10.3390/COMPUTERS12080150.

[13] M. S. A. Abrar, S. K. Nishat, S. M. Muhammad, et al., "Deep Learning-Based Motion Style Transfer Tools, Techniques and Future Challenges," Sensors, vol. 23, no. 5, pp. 2597–2597, 2023, doi: 10.3390/S23052597.

[14] K. Ryotaro and K. Seiichiro, "A generative model of calligraphy based on image and human motion," Precision Engineering, vol. 77, pp. 340–348, 2022, doi: 10.1016/J.PRECISIONENG.2022.06.006.

[15] B. Anthony, G. RyanRhys, G. Robert, et al., "Generative model-enhanced human motion prediction," Applied AI Letters, vol. 3, no. 2, pp. e63–e63, 2022, doi: 10.1002/AIL2.63.

[16] B. Q. Zhao, Y. Y. Fu, Z. Su, et al., "Review on 3D digital human motion generation guided by multimodal information," Journal of Image and Graphics, vol. 29, no. 9, pp. 2541–2565, 2024.